DOCUMENT RESUME

ED 165 117 CS 004 614

AUTHOR Wardrop, James L.; And Others
TITLE A Framework for Analyzing Reading Test Characteristics. Technical Report No. 109.
INSTITUTION Bolt, Beranek and Newman, Inc., Cambridge, Mass.; Illinois Univ., Urbana. Center for the Study of Reading.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Dec 78
CONTRACT 400-76-0116
NOTE 65p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Predictive Validity; *Reading Comprehension; *Reading Research; *Reading Tests; *Test Construction; *Testing; Test Selection
IDENTIFIERS *Center for the Study of Reading (Illinois)

ABSTRACT

To provide a structure for describing different approaches to testing, five dimensions along which reading tests may differ have been identified. They are: test uses, item generation, item revision, assessment of precision, and validation. Each dimension contains variations that reflect O.K. Buros's distinction between differentiation and measurement. These dimensions were used to profile representative reading comprehension tests. Only standardized, norm-referenced achievement tests, whose uses emphasize differentiation, were found to have an inference system consistent with those intended uses. No tests were found having inference systems consistent with such intended uses as certifying competence, diagnosing strengths and weaknesses, and tracking progress--uses that emphasize the measurement function of tests. Tests constructed using a domain-referenced approach would yield such an inference system and fill some gaps in the array of currently available measures of reading comprehension. (Author/FL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED165117

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CENTER FOR THE STUDY OF READING

Technical Report No. 109

A FRAMEWORK FOR ANALYZING READING TEST CHARACTERISTICS

James L. Wardrop, Thomas H. Anderson,
Wells Hively, Richard I. Anderson,
C. Nicholas Hastings, and Keith E. Muller

University of Illinois at Urbana-Champaign

December, 1978

University of Illinois
at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

The research reported herein was supported in part by the National Institute of Education under Contract No. US-NIE-C-400-76-0116.

All authors are at the University of Illinois at Urbana-Champaign except for Wells Hively, who is at the Spaulding Youth Center, Tilton, New Hampshire, and Keith E. Muller, who is now at the University of North Carolina, Chapel Hill.

CS004614

Abstract

In order to provide a structure for describing different approaches to testing, five dimensions along which tests may differ are identified. The dimensions are (1) test uses, (2) item generation, (3) item revision, (4) assessment of precision, and (5) validation. Within each dimension, variations are described reflecting Buros' (1977) distinction between differentiation and measurement. These dimensions are used to profile representative tests from the area of reading comprehension. Only standardized, norm-referenced achievement tests, whose uses (dimension 1) emphasize differentiation, were found to have an inference system (dimension 2 through 5) consistent with those intended uses. No tests were found having inference systems consistent with such intended uses as certifying competence, diagnosing strengths and weaknesses, and tracking progress--uses which emphasize the measurement function of tests. Tests constructed using a domain-referenced approach would yield such an inference system and fill some gaps in the array of currently available measures of reading comprehension.

A Framework for Analyzing Reading Test Characteristics

In the last few years, labels describing various forms of testing have proliferated. Norm-referenced tests, criterion-referenced tests, objectives-referenced tests, domain-referenced tests, and others have been described and discussed as if each somehow represented a unique form of assessment. One of the results of this proliferation has been, predictably, confusion.

One reason for the confusion has been a failure to differentiate among the various dimensions along which differences are said to exist. For example, the distinction between norm-referenced and criterion-referenced measurement is, in the first place, a distinction between two kinds of score interpretation, although one's choice in this respect affects other characteristics of the test content and the development procedures (cf. Wardrop, 1976, ch.8). On the other hand, the distinction between objectives-referenced and domain-referenced measurement reflects the difference between two approaches to item development, although one's choice here has implications for subsequent scoring and interpretation procedures.

The proliferation of labels, procedures, and claims derives from a preoccupation with the specific settings and purposes for which tests are created. Labels have been used to distinguish among products (i.e., tests) rather than among the fundamental choices made--often implicitly--in designing and creating a test. This paper takes a different approach. It is intended to provide a structure for understanding the unique position of domain-referenced testing within the total array of

models and techniques for measuring achievement. The conceptualization presented makes explicit several of the dimensions along which various kinds of tests may (and do) differ.

The first section contrasts what might be termed "idealized forms" of typical standardized achievement tests (designed to differentiate among examinees) and domain-referenced measures (designed to measure examinee status), because these two approaches closely approximate extreme positions on the dimensions used. In section two, the dimensions representing (1) uses of tests, (2) item generation paradigms, (3) item revision procedures, (4) approaches to assessing the precision of measures, and (5) test validation techniques are elaborated and described in some detail. Section three demonstrates how these five dimensions can be used to characterize ("profile") some of the major types of reading comprehension tests. This demonstration is accomplished by actually locating each type of test on each dimension, thereby demonstrating some of the similarities and differences among the tests. Finally, in the concluding section the issue is addressed of how logical contingencies across these proposed dimensions lead to inferences about appropriate and inappropriate uses of the various kinds of measures, emphasizing the important role of domain-referenced measures in the total context of procedures for assessing reading comprehension.

Standardized Achievement Tests and Domain-Referenced Measures

Differences in intended uses of test results have implications for what test characteristics are most salient, for how test items are created and revised, and for how one proceeds to judge the quality of the tests. In

this section, two extremes on a continuum of uses are proposed, and some implications of those uses are identified.

Differentiating or Measuring: The Distinction and its Implications

Most simply, tests are used either to differentiate or to measure (Buros, 1977). Differentiation typically is the basis for assignment of examinees to categories (hired-rejected; admitted-denied; gifted-normal), that is, for making selection decisions when access is limited. (Rank ordering permits such decisions to be made even if the numerical location of category boundaries is changed.)

Measurement--estimating the quantity of a characteristic--typically leads to decisions about each individual in a situation where access may be, at least in principle, open to everyone (pass-fail; certified-not certified; given new instructional material-recycled through previous material; assigned activity x-assigned activity y). Here the purpose is not to choose some among many but to provide guidance to each individual (cf. Buros, 1977).

Intended uses of test results affect the kind of test required. A test used to rank order or classify is typically norm-referenced in the sense that an individual examinee's score acquires meaning primarily through its relationship to the scores of other examinees. In this case, the importance of the reference group (for score interpretation) is highlighted. In contrast, a test used to assess status or change on some scale of measurement depends for its utility on the soundness of the theories, rules, and content that underlie the instrument and guide the production of test items. In this case, the importance of the quantity of the characteristic being measured is highlighted.

A test used to differentiate need only be administered once (or at infrequent intervals), since the rank ordering of individuals on the traits underlying it is thought to be relatively stable over time. On the other hand, a test used to measure status or change should exist in such form as to permit frequent administration so that instructors and students can see the change take place. In the former case, only a few standard forms of the test are required; in the latter, it becomes necessary to have available many alternate versions of the test.

Creating Tests for Differentiating or Measuring

The different potential uses for and requirements of tests imply different models which govern the production of test content. Tests for differentiating tend to be based on descriptive models which usually take the form of a table of specifications created by crossing process objectives with content categories to form a rectangular grid. Weights reflecting expert judgment about the importance of each skill-content combination are assigned to the cells within this table of specifications to indicate the proportion of test items that should represent each cell in the grid.

Tests for measuring quantity are based on predictive models of the behavioral domain in which assessment is to take place. Such models focus on the specification of variables affecting item difficulty, transfer, and generalization. With such a model, it becomes possible to predict in advance the relative difficulty of items for an examinee.

Although the models that guide the selection of test content provide different frameworks within which to work, subsequent procedures for item (or item form) writing and for trial administration of items are relatively

standard, regardless of the intended use of the test. There are differences, though, in how the information obtained through trial administration is used. For tests used to differentiate, the focus is on item-test relationships, and items are revised or replaced in order to maximize those correlations. With measurement instruments, on the other hand, it is not particular items but generic item forms that are tried out. Where the focus is on predicted performance, it is necessary to revise the theory (model of behavior) underlying the development of the instrument, rather than merely revising or discarding individual items (Anderson, Wardrop, Hively, Muller, Anderson, Hastings, & Frederiksen, 1978).

Determining Test Quality

The different purposes for which tests are used also suggest differences in how one goes about evaluating the quality--i.e., reliability and validity--of the instruments that are created. Tests for differentiating are reliable to the extent that they consistently yield the same ordering on repeated administrations. The various traditional approaches to estimating reliability provide good indicators of this sort of consistency (cf. Lord & Novick, 1968). For measurement instruments, the concept that best approximates reliability might be termed "repeatability." Such measures will, in the absence of change in the attribute being assessed, yield consistent estimates of the status of an individual over repeated occasions. Alternatively, they will permit the fitting of appropriate mathematical functions to a pattern of change over time. Figure 1 illustrates these notions of repeatability.

Insert Figure 1 about here.

For estimating the validity of a test, again the two types of instruments require different approaches. Evidence for the validity of tests used for differentiating is most often given in two forms: content validity and predictive validity. Content validity reflects the adequacy of the sampling of cells in the process-by-content grid; and predictive validity indicates the degree to which the rank ordering of examinees on the basis of test performance agrees with the rank ordering in terms of some other criterion, that is, the extent to which test performance predicts criterion performance (cf. Cronbach, 1971). The validity of a measuring instrument is supported by demonstrating its adequacy in representing the model from which it was developed and by verifying predictions (made on the basis of that underlying model) about the relative difficulty of items and item forms. This notion of validity is closely related to some aspects of construct validity (Cronbach & Meehl, 1954).

By way of summarizing the distinctions made above, consider the two descriptions shown in Table 1. The material presented thus far has been

---

Insert Table 1 about here.

---

couched in language that suggests sharp contrasts among the features of two types of test. The more complete treatment that follows extends these contrasts into a larger context.

A Descriptive Model of Tests and Measures

Several features of tests have been identified which distinguish between norm-referenced achievement tests and domain-referenced measures. Five of these features are now described in greater detail. This treatment

leads to the generation of a "first order model" with which test instruments can be described and evaluated. Figure 2 shows the framework of the model.

Insert Figure 2 about here.

Uses of Tests

Major uses of tests can be clustered into two groups, ranging from a focus on relative selection (differentiation) to a focus on absolute assessment (measurement). Uses which emphasize relative selection include (a) maximizing job performance, where "job" includes the job of learning in an instructional setting; (b) insuring fairness in the allocation of educational or vocational positions; and (c) minimizing effort and disappointment in training. Uses which emphasize absolute assessment include (a) certifying competence, (b) diagnosing strengths and weaknesses, and (c) tracking progress.

Uses emphasizing relative selection can be differentiated in terms of who, from an economic point of view, is the consumer of the test results. When tests are used to maximize job performance, the consumer is the institution or agency that has a job to be done. A classic example has been pilot training in wartime, where the objective is to minimize training time and failure rate. Job performance is the criterion, and the testing problem is to provide inexpensive ways to select candidates having the best chance of success from a large pool of applicants.

When the task is insuring fairness in job allocation, one may think of the consumer as being society in general. Given that desired programs of professional training or skilled occupations can accommodate only a

limited number of individuals, the available places should be allocated to those persons most likely to do well in them, regardless of influence or social position. This is the sociological reason behind college entrance examination boards, scholastic aptitude tests, etc. Performing this selection function is one of the major purposes of public education, and a primary use of standardized tests at all levels is gradually to sort and "track" youngsters through the system into "appropriate" places in trades and professions.

Minimizing effort and disappointment in training is a use of tests which may be thought of as the mirror image of fairness in job allocation. Here the consumer is the individual, and the objective is to collect information that helps individuals choose among alternative professions or training programs on the basis of information about their probability of success. In good high school counseling programs, for example, it is possible for a youngster to consult tables of probability of success in a given college program for students with particular high school grade averages and scores on college entrance examinations. Here, the focus is on predicting personal success in competition with others.

At the other end of the scale are those uses of tests which emphasize absolute assessment. Two closely related uses are to diagnose strengths and weaknesses and to track progress. The consumers are members of the team consisting of teacher and student. Together they need to know in what areas of knowledge and skill the student needs to work, and they need to keep track of progress as these areas are mastered. An interrelated, but slightly different, use of tests is to monitor change for groups of students. In this instance, the consumers may be a group of educational

policy makers. Such an orientation is essentially the motivation for state and national assessment programs. Group, rather than individual, performance is the focus, but the processes are the same.

Finally, when tests are used for certifying competence, the consumer is society and the problem is making sure that individuals who perform essential jobs have the skills necessary to do them reasonably well. The goal is to protect the public, and the tests may be thought of as licensing examinations. Tests of minimal competencies in basic education fall into this category.

An important point to keep in mind, and the basic theme throughout this paper, is that certain uses of tests imply certain test construction procedures and mechanics of inference. Test purposes emphasizing differentiation (toward the left side of the uses scale in Figure 2) imply the mechanics of inference associated with norm-referenced testing, which run down the same side of the four subsequent scales. Test purposes emphasizing measurement (toward the right-hand side of the uses scale in Figure 2) imply the mechanics of domain-referenced testing, which run down the same side of the other scales. One cannot effectively change the purpose of a test from differentiation to measurement, or attempt to use a test for more than one purpose, unless the corresponding mechanics of inference are all available. As is shown later, it is very instructive to analyze the properties of currently available tests by profiling them along the various scales.

Item Generation

The item generation dimension includes four identifiable points which range between end-points labeled "descriptive categories" and "generative

rules." "Descriptive categories" identifies the basis from which norm-referenced test items are typically generated; "generative rules" identifies the basis from which domain-referenced test items are generated.

The use of a table of specifications is located at the "descriptive categories" end of this scale. Gronlund (1976) illustrates several tables of specifications where objectives form one dimension in the table and content areas form the second dimension. A cell in the table (the intersection of an objective with a content area) can be used as an idea source for potential test items. For example, 'knows specific facts' about 'city life' is an objectives-by-content-area cell in a table of specifications for a third grade social studies test.

A weighting system, whereby each cell in the table is assigned a numerical value, often is used to help determine the composition of items that will be organized into a test. Using the above example, the cell (knows specific facts about city life) is not as important as the cell (understands principles and generalizations about city life). Consequently, a larger proportion of test items would be included about the latter than about the former.

In summary, a table of specifications is formed by making a double-entry grid of objectives and content areas. Each cell in the grid is given a weight which represents the proportional contribution of that cell to the collection of test items. Item writers and test developers consult this grid to insure that the final test is consistent with its intended specification.

Another descriptive approach to item generation involves the use of a list of objectives without theory. This technique begins with the

construction of an extensive list of performance objectives. Each objective is entered into the list, not on the basis of some well-developed theory which defines the area to be assessed, but rather because the objective is thought to be an important outcome in that area. There are seldom explicit rules for generating items from each objective, but it is commonly thought (e.g., Anderson & Faust, 1973) that properly stated performance objectives can be easily and directly converted into test items. In this case, the degree of ease and directness with which the test items can be generated depends upon how well the objectives were constructed.

A related point on the scale, but one requiring greater precision, is labeled "ordered list of objectives." Objectives in an ordered list are those which not only precisely define the intended behavior, but also show some relationship among the objectives. For example, performance objectives in mathematics are sometimes seen in ordered sets or skill hierarchies. It is usually assumed that before one can learn to add two-digit numbers one must learn to add one-digit numbers.

The point labeled "theoretical partitioning of a specified set" most closely approximates the end-point, "generative rules." In this scheme, rules or algorithms are explicitly outlined to prescribe the entire domain of items. Each rule or algorithm is derived from a specific theory which fully partitions the area to be assessed. Ambiguity is minimal when constructing or selecting items because all items are the products of explicit generative rules. Martuza (1977) presents a number of examples of tests that are created using item generative rules. Two of the four approaches are discussed below to characterize this approach to item generation.

Martuza's linguistic transformation approach is illustrated by efforts of Bormuth (1970) and Anderson (1972). A simple transformation rule under Bormuth's scheme requires that a base sentence be transformed into an item by replacing the period at the end of the sentence with a question mark (e.g., "Jane picked up the ball." becomes "Jane picked up the ball?"): the item is then answered with "yes" or "no." A more elaborate scheme suggested by Bormuth requires the selection of a presumably important sentence or phrase followed by a conversion of it into a "wh-question": a question beginning with, for example, who, what, or when. Specific rules are provided to guide the item writer in the selection of the most appropriate wh-word.

The scheme proposed by Anderson (1972) is based on the proposition that to measure comprehension achievement, one must ensure that the students cannot answer questions on a test simply by dealing with the surface form of the message. If what is to be measured is the comprehension of a concept, principle, or rule, the student should be asked either to recognize or to generate an example of that concept, principle, or rule--one that the text does not present. If the material to be learned is not a concept, principle, or rule, but perhaps a fact or a statement of relationship between several facts, then the statement should be written in a paraphrase form and the student should be asked to recall, or perhaps generate, essential deleted parts of that statement. In either case, whether by example or by paraphrase, the student must have processed the written message in a deeper form than a mere surface interpretation in order to answer the question correctly.

A second approach labeled "item forms" is described by Hively, Maxwell, Rabehl, Sension, and Lundin (1973). In this technique the syntactic structure

of several items is similar, with differences due to the replacement of specified components of the structure with specified alternatives. Selective choice of the alternatives that complete the item structure allows one to define completely a simple, straightforward domain.

It is important to note that there can be several sets of rules operating simultaneously during item generation. This is especially true for reading comprehension items. The typical format of reading comprehension tests is to have students read a passage and then answer questions about the passage. Experience with developing these kinds of items has led to the conclusion that it is not only necessary that a set of rules be used to generate the questions (that is, the probes that elicit a preferred set of behaviors from students), but it is also important to use a set of rules to generate or select the passages for students to read and process and to generate the answer alternatives (if using a multiple-choice format) or the procedure for scoring (if using other response formats).

Item Revision

The item revision dimension is described at one end by the process of selecting and adjusting test items and at the other end by modifying rules for selecting and generating items. For tests used to differentiate, the item generation phase results in a large number of items, perhaps as many as two or four times the number of items in the proposed test. After some preliminary editing, the items remaining in the pool are field tested, i.e., administered under test conditions to students of the target population. Then at least two psychometric indices are computed for each item from the field data. These are a discrimination index (whether or not each item

discriminates well between those students who are able to perform this skill adequately and those students who perform it inadequately), and a difficulty index (the proportion of students answering the item correctly). The items are either revised by staff item writers or contracted to outside consultants for revision. At this point the selection of an item is determined primarily by its psychometric features, although the guidelines for test content, whether in the form of a table of specifications or a list of objectives, are also used to ensure that the distribution of items across categories remains consistent with the a priori specifications.

With repeated revision, however, individual items may take on the characteristics of several objectives or cells in the table of specifications. They may even stray from one category to another. Items are continually being changed, modified, and updated so that the end result may be a set of items that does not represent the coverage originally intended.

In the middle of this scale is the process for item revision called "adjusting objectives." When several items have been developed to measure a particular objective, it sometimes happens that the performance of examinees is inconsistent from one item to another. Sometimes, inspection of items and the objective reveals that the objective is vaguely worded. When this happens, revising the objective leads to the elimination or revision of some items. In other instances, the objective is found to be too complex, subsuming several skills or concepts that are psychologically distinct. In these cases, revising the objectives involves dividing them into a set of more narrowly-focused objectives. Items themselves may be simply sorted into the new categories created by this process, but otherwise left unchanged.

In other instances, when tests have been developed using an ordered list of objectives, examinee performance may suggest that some objectives are misplaced in the list (cf. Macready, 1975). When such anomalies are identified, the list of objectives may be re-ordered to correspond to the empirically-determined sequence.

The danger that items will be so altered during the revision process that they no longer fit the categories for which they were originally written can be minimized by employing the procedure indicated at the measurement end of the dimension, i.e., establishing rules for selecting and generating items. In this case, when so-called problem items are identified, the assumption is that the problem is not primarily with the items but with the rules. Accordingly, the preferred procedure is not to modify the items, but to modify the rules. For example, when generating a set of passages to be used in reading comprehension tests, writers may be operating under a rule which states "the more syntactically complex the sentences, the more difficult the passage will be to comprehend." A rule for generating syntactically complex sentences may be "to use sentences of fifteen words or more." If this rule is followed and if other variables are held constant, one would expect students who were given items with long sentences to perform less well on a probe about comprehension than students who received paragraphs with short sentences. If the data are not consistent with this expectation, the point of revision would be the rule and sometimes the theory used to generate sentences, and not the items that ask students what the paragraph is about. It may be necessary in the above case to use more elaborate rules about syntactic complexity.

Assessment of Precision

The precision with which a test measures what it is designed to measure is an essential characteristic for judging the quality of that test. The scale along which tests vary in terms of how that precision is estimated has at its endpoints (a) procedures based on intersubject variability (classical conceptualizations of reliability) and (b) procedures based on intrasubject consistency (what we have chosen to call "repeatability"). Intermediate between these two extremes is generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

All standardized, norm-referenced achievement tests and most other instruments utilize procedures based on intersubject variability to obtain estimates of precision (reliability). The exact procedures differ somewhat from test to test, but essentially the goal is to estimate the reliability with which the test will consistently differentiate among persons, i.e., the consistency with which examinees will be rank-ordered on repeated (and, at least hypothetically, independent) measurements with the same test form. Procedures using retesting (after some reasonable time has elapsed), split halves, or internal consistency (coefficient alpha) all seek to arrive at an estimate of the reliability of a single test form. For many tests, two or a small number of alternate forms are developed and the correlation between examinee performances on these alternate forms serves as an estimate of the reliability of each form. This procedure is also tied to the specific test forms used and is based on the notion of replicable rank ordering of examinees.

As the focus of concern shifts from a small number of forms of a test to test forms as samples of behavior in some domain, classical notions of

precision, based as they are on a belief in statistically parallel measures as the basis for psychometrics, become inadequate. Under this alternative focus, a specific test form is of interest only insofar as it represents some collection of what Lord and Novick (1968, Ch. 8) refers to as "nominally parallel" measures: measures whose content is obtained through some representative (random or stratified random) sampling from a universe of behaviors reflecting the trait of interest. The similarity of this concept to that of a domain-referenced measure is apparent.

The classical approach to test reliability focuses on the particular test form at hand. The concern is with estimating how dependably one can rank order examinees on the basis of their performance on that particular test. Generalizability theory, on the other hand, takes as the primary datum of interest an inferred "universe score" for an individual. Any particular test form is of interest only insofar as performance on that test provides a basis for inferring how examinees would perform if administered all items in the universe being sampled. Since the particular test form is viewed only as one of many possible representations of behaviors in the universe, a "generalizability coefficient" is obtained that indicates the precision with which the universe score can be estimated.

It is important to recognize that in the use of nominally parallel test forms, the emphasis is still on rank ordering with respect to the characteristic being assessed. Accordingly, although it accommodates explicitly the complexity introduced by allowing for test forms that are not statistically parallel, generalizability theory is still based on what is essentially an "intersubject variability" approach to the estimation of

precision. If one is using a domain-referenced approach to testing and wishes to use results from such testing to discriminate among examinees, one cannot do better than to employ the methodology of generalizability theory to assess the precision of the measures.

When one's interest is not in assessing relative status of individuals, but in making statements about the level of attainment--or change in level of attainment--of an individual over time, neither classical reliability theory nor generalizability theory is sufficient. This shift in emphasis from relative standing to individual status or change with respect to a domain requires that precision be conceptualized in terms of replicability of outcome over repeated measures, if the characteristic is stable. Precision must be thought of in terms of the existence of (relative) regularity and smoothness of growth curves, if the characteristic changes over time for a particular examinee. Techniques appropriate for this kind of assessment of precision include time series analysis and function fitting.

Implicit in this last notion of precision is the existence of a rather large number of test forms, all representing the same characteristic. In other words, the approach presumes both a domain-referenced strategy for test construction and a focus, in score interpretation, on the temporal pattern of performance of an individual examinee.

Validation Procedures

For this analysis, five types of validity have been identified. For tests used to differentiate, predictive and content validity are most appropriate, while for tests used to measure, considerations of construct and structural validity are paramount.

The predictive validity coefficient is simply the correlation between the test, taken as a predictor, and some other measure taken as the criterion. A closely related idea is that of concurrent validity, where the task is to predict performance on an "expensive" test with a less costly one. In that case, a group of subjects is given both tests, and the scores on the more expensive test serve as a criterion. With these methods, validity is defined with respect to performance on some criterion external to the test (Lord & Novick, 1968, Chapter 12).

The second point on the scale is content validity. An educational achievement test may be said to have content validity to the extent that it measures those goals which are the objectives of instruction in the area. Although content validity has rarely been held in the same esteem as predictive validity, it must be present either in the test itself or in the criterion measure. At the same time, content validity is probably the type most easily understood. One must be careful to recognize tests which relate directly to the materials of instruction but not to the objectives. A test could spuriously be said to have content validity if it sampled effectively from each of the instructional behaviors but did not assess any of the target behaviors. The use of a table of specifications is one heuristic means of attempting to ensure content validity.

Face validity is a concept with limited value. A major problem has been the ambiguity in the definition of the term. In all cases an appeal to "common sense" is made, and the "appearance" of validity is emphasized. This emphasis on appearance rather than fact limits the utility of face validity.

Although the history of construct validity is shorter than that of the other types, a rather large body of literature has appeared on this topic. A construct is usually thought of as a hypothetical attribute of persons, an attribute that will presumably affect test performance. Construct validation is the process of determining whether performance on a test is consistent with predictions derived from an elaborated theory about the construct. The procedures for establishing a test's construct validity are many and varied, including both correlational and experimental studies. The analysis of intertest relationships, the effectiveness of experimental interventions, the confirmation of predictions about differences in test performance among groups known to differ with respect to the construct: these are but a few of the techniques that are employed to assess a test's construct validity (see Cronbach & Meehl, 1955).

Structural validity is a term introduced by Loevinger (1957), drawing on the earlier work of Coombs (1953) and Peak (1953). According to Loevinger, structural validity ". . . refers to the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured" (p. 660). Although she uses the term to include both non-test manifestations of the trait and the degree of inter-item structure, the latter component is emphasized in the discussion.

Closely related to construct validity, structural validity applies to tests that have clearly defined procedures for creating items. If the structure of a test is derived from some model of the behavior to be assessed, the structure itself should reflect those features of a task presumed to affect the difficulty of the task. In addition, structural relations within

a test may suggest what strategies could be used to perform the task successfully. In such a situation, it is possible to infer from the test structure what specific outcomes should occur in any research in which the variables used to create the test are manipulated. Conversely, and more conventionally, knowledge about the interrelationships among non-test manifestations of performance in a domain may lead to predictions about the interrelationships among responses to items from a test referenced to that domain.

For example, a reader's ability to find the main point of a passage might be affected by a feature of the text such as the frequency of main point statements. If repeated statements of the main point make it easier for readers to comprehend what the passage is about, one should find in a testing situation that a greater proportion of examinees are able to state the main point of a passage containing such repetitions than of a passage containing only one such statement. To the extent that the relative difficulty of test items is consistent with non-test performance, the test may be said to have structural validity. Alternatively, if a test is designed such that its items are intended to measure ability to identify, say, four different kinds of text distortion, a factor analysis of test responses might lead to the identification of four factors underlying the responses. Again, this result would be evidence for the structural validity of the test.

Profiling Major Types of Tests on Major Dimensions

The five dimensions described in the previous section provide a basis for comparison of the various types of tests most frequently used. More specifically, these dimensions can be used to create "profiles" of (a) standardized achievement tests, (b) criterion-referenced tests, (c) individual

psycho-educational tests, (d) curriculum-associated tests, and (e) national assessment instruments. In this section, all five of these test types are profiled, using tests of reading comprehension as examples.

Standardized Achievement Tests of Reading Comprehension

The standardized tests that were included in the Anchor Test Study (Biancini & Loret, 1974)--the California Achievement Tests (CAT), Comprehensive Tests of Basic Skills (CTBS), Gates-MacGinitie Reading Test (G-M), Iowa Tests of Basic Skills (ITBS), Metropolitan Achievement Tests (MAT), Sequential Tests of Educational Progress (STEP), SRA Achievement Series (SRA), and Stanford Achievement Tests (SAT)--account for nearly all the standardized, norm-referenced achievement tests commonly used in the schools. (The Gates-MacGinitie Reading Test differs in several respects from the other seven tests named above. Because so little is reported about its development and technical characteristics, it is excluded from the following discussion.)

For all of these tests, the primary use is to provide an indication of general achievement in reading. Performance on these tests is given meaning by considering an examinee's performance relative to that of others who have taken the test (a norm group). Raw scores may be converted to grade equivalents, percentiles, stanines, or some sort of standard score. These converted scores all indicate the examinee's status with respect to others. Thus, they are all fundamentally indicators of rank order.

If scores from these tests are used to make instructional decisions, the purpose is either to group examinees who perform at similar levels (i.e., "ability grouping"), so that they may be given common instruction,

or it is to identify performers who are extremely deviant from the group and who may need special individual attention. In either case, decisions are made on the basis of pupils' relative standing on the test. In brief, justifiable uses for these tests all focus on relative selection, i.e., differentiation.

One of two item generation approaches is typical: to create a table of specifications (e.g., SRA, STEP, ITBS) or to work from a list of instructional objectives (e.g., SAT, MAT). Teams of item writers, either professional educators hired as consultants or in-house employees of the test publisher, are given the specifications or objective: typically some rudimentary guidelines for item generation are also provided. These teams then produce a large pool of items--usually two to four times as many as are to appear in the final test form--to be tried out.

Formal item revision takes place after experimental versions of the tests are produced and administered to a representative sample of the target population.

> From this tryout, the publisher will obtain information about item-difficulty levels (the proportion of pupils answering each item correctly), how well the individual items discriminate high-scoring from low-scoring pupils on the total test (a good item being one that is more frequently answered correctly by pupils in the former group than by those in the latter), and -- for items administered at several grade levels--the extent to which progressively greater percentages of pupils answer the item correctly as grade level increases. Items that are too easy or too hard, items that do not discriminate between high- and low-scoring pupils, and items for which the grade progression is inappropriate are [revised or] eliminated. (Wardrop, 1976, p.73)

The precision (reliability) of these tests is reported using the statistics of intersubject variability: internal consistency (split-half or coefficient alpha), alternate forms, and--less frequently--test-retest estimates. These tests generally have two alternate forms, so that repeated measurement of individuals is infeasible; and they do not consist of items that can be treated as samples from some universe, so that coefficients of generalizability are inappropriate.

For standardized achievement tests, the major validity claim is based on the adequacy with which their content represents that of the major curriculum series and other reading material commonly encountered in the schools. Other evidence for validity may be adduced in the form of correlation coefficients relating achievement test scores to aptitude measures or to other indicators of achievement. An implicit claim is that the ordering of pupils on the basis of test performance is consistent with their subsequent academic performance. In practice, the evidence suggests mainly that the ordering of pupils on one achievement test form is consistent with the ordering obtained when they are administered another achievement test form at some later time.

This analysis of standardized tests of reading comprehension yields the profile that appears in Figure 3.

Insert Figure 3 about here.

Criterion-Referenced Measures: SOBAR READING

Our survey of commercially available materials turned up two large scale criterion-referenced programs, Houghton-Mifflin's SCORE program and

the Science Research Associates (SRA) SOBAR program. The two programs are very similar and may be reasonably thought of as prototypes which illustrate the state of the art in the commercial field of criterion-referenced testing. Since the SOBAR program is more thoroughly developed, it is used for analysis.

The SOBAR testing system is marketed to school districts. It offers the following sets of objectives for reading: phonics analysis (32 objectives), structural analysis (43 objectives), vocabulary (13 objectives), comprehension (25 objectives), and study skills (27 objectives). The school district's curriculum committee may choose any set of objectives from the list to match the objectives which are emphasized in the district's instructional program. SRA then provides tests by assembling items for the chosen objectives from their item bank. If the district does not wish to choose objectives, SRA has a standard set of tests which sample items from the objectives in such a way as to provide what SRA believes to be a good general representation of the objectives. Three items are generated for each objective.

The tests are administered by the district and scored by SRA. Results are presented in a variety of forms, including individual student profiles (emphasizing the performance of an individual student on all of the objectives), and list reports (summarizing the interactions between student performance and difficulty of objectives across a class, school, or district).

The major use of SOBAR is to aid in diagnosing strengths and weaknesses. Test results seem useful to teachers for making changes in classroom instruction and to school administrative groups for analyzing overall curricula. As a supplement to the system, SRA offers further diagnostic testing keyed

to each of the objectives, with cross references to teaching materials in major reading curricula.

Since it takes approximately seven weeks to get test results following submission of tests for scoring, tracking an individual's progress is not a major use. Certification of competence is not emphasized, but it certainly can be approached with the system.

Item generation appears to have involved procedures somewhere between a list of objectives without theory and an ordered list of objectives. Items may have been tried out and revised or discarded if they proved to be unexpectedly difficult.

None of the procedures for assessing precision seems to have been used. Variability of performance within and between objectives is the general issue, and it appears not to have been investigated. Data are needed concerning the probability of success for an individual on randomly selected items related to a specific objective, given that individual's pattern of hits and misses on preceding items. In the absence of such data, it may be very misleading to suggest a mastery criterion of three correct out of three attempted, although this is certainly better than making inferences about mastery from an individual's response to a single item.

Validation seems to be a matter of expert judgment of content. There have been no studies of construct validity based on patterns of students' performance within and across objectives. Furthermore, no correlation of overall performance with general measures of reading ability are reported, although these would no doubt be high.

The overall profile of the SOBAR system is depicted in Figure 4.

Insert Figure 4 about here.

Individual Psychoeducational Tests: Illinois Test of Psycholinguistic Abilities

The Illinois Test of Psycholinguistic Abilities (ITPA) is an example of a group of tests called "psychoeducational" tests. Tests in this group are prepared especially for individual administration to children thought to have learning disabilities. According to Anastasi (1976), such tests should be considered as observational aids for the clinical psychologist and learning disabilities specialist. Thus, while not reading tests per se, they are used to help diagnose possible causes of a child's reading problems. They are similar to criterion-referenced tests in their attempt to portray the child's strengths and deficiencies in narrowly defined domains. However, scores in the domains are norm-referenced, because decisions about whether a performance is deficient are based on data from "normal" children.

An experimental edition of the ITPA was first published by McCarthy and Kirk (1961). This edition contained nine subtests which attempted to tap educationally-important abilities as defined by Osgood's (1953,1957a, 1957b) theoretical model of communication processes. In 1968 the test was revised "to improve the subtests of the battery and to add tests which were not included in the original battery" (Paraskevopoulos & Kirk, 1969, p. 10). This version, which is profiled below, includes twelve subtests. The administration of two of the subtests is considered optional.

The authors of the ITPA claim that it is a diagnostic test for cognitive dysfunction and learning disabilities in children two through ten years of

age. "Its object is to delineate specific abilities and disabilities in children (in the area of learning and communication) in order that remediation may be undertaken when needed" (Kirk, McCarthy, & Kirk, 1969, p. 5). In order to achieve this goal, each subtest was normed by age level against a population of "average" children. In this manner a standard of competence is defined which can, in a sense, be used as a criterion against which to judge the presence of a specific disability. Thus, uses of the ITPA range from certifying competence to diagnosing strength and weaknesses.

Osgood's (1953, 1957a, 1957b) communication process theory, on which the ITPA is based, contains three dimensions: (1) channels of communication, (2) psycholinguistic processes for dealing with information, and (3) levels of organization of the information. Each subtest was constructed to represent the intersection of one and only one aspect of each of the three dimensions. For example, the first subtest is of Auditory Reception and involves the auditory-vocal channel, the receptive process, and the representational level. To avoid either tapping the expressive process or measuring the child's ability to obtain meaning from syntax, items were generated in the format, "Do (Noun) (Verb)?" In this way, only a "yes-no" response, which the authors claim needs only minimal expressive processing, is required. Syntax is constant regardless of the difficulty of the item. In brief, item generation involved the use of theory-based generative rules.

Before the final test was assembled, large pools of items were administered to a sample of children to provide stable statistical information. Item analyses provided information about item homogeneity, age differentiation, difficulty level, and discriminating power. This information guided item revision although some adjustment of the original theory occurred.

The precision of the ITPA was estimated by using indices of internal consistency (K-R 20 and Hoyt's ANOVA), test-retest stability, reliability of differences between subtest scores, and standard error of measurement. Also, interscorer reliability was examined on one of the subtests (Verbal Expression).

Unfortunately, little work has been done towards validating the revised version of the test. Indeed, the test has been criticized by several authors (e.g., Carroll, 1972; Chase, 1972) for the lack of validity information. McCarthy and Olson (1964) evaluated the 1961 edition using measures of concurrent and predictive validity, content validity, face validity, and construct validity--procedures which encompass quite a large part of the validation dimension. However, Carroll (1972) notes that results of several studies of construct validation for the test have been equivocal.

Figure 5 is a graphic summary of this analysis of the ITPA. Other indiviudal psychoeducational tests would certainly lead to different profiles,

---

Insert Figure 5 about here.

---

but the ITPA seems as good a representative of this category of tests as any that might have been used.

Curriculum-Associated Tests

A large number of tests either accompany or are an integral part of reading curricula. This section does not focus on any specific curriculum, although specific examples are used for purposes of illustration. The following analysis is based upon information obtained from the literature associated with curricula, from interviews of curriculum developers/

publishers, and from Jenkins and Pany's (1977) "Reading Comprehension in the Middle Grades: Instruction and Research."

A number of curricula include tests that are claimed to be useful either for diagnosing strengths and weaknesses and/or for tracking student progress. Many of these tests are labeled as criterion-referenced, a characteristic that some developers/publishers equate with the ability to function in these ways. Aspects of the Wisconsin Design for Reading Skill Development (WDRSD)are described below to illustrate some of these points. The WDRSD includes criterion-referenced tests, the Wisconsin Tests of Reading Skill Development (WTRSD) keyed to single objectives of the curriculum's hierarchy of objectives and existing in two parallel forms (Otto & Askov, 1974). One form of each test at a selected hierarchy level is administered to each student in order to diagnose strengths and weaknesses across objectives. Students who perform similarly are grouped together for instructional purposes. Results of each test indicate little about which strengths or weaknesses exist within an objective; instead, they identify competence, defined by an 80% criterion level, at a skill portrayed by an objective.

"Guides to Individual Skill Assessment" are offered to instructors to help them monitor the progress of students towards skill attainment. Once an instructor feels a student has attained a particular skill, the second form of the test for that skill is administered. If a student fails to achieve the criterion level on this form, more instruction is warranted, and subsequent decisions as to skill attainment must be made informally. If a student achieves the criterion level on this second form, attention is directed to other skills at the same hierarchy level that have yet to be

mastered, or the WTRSD for the next level of objective is administered to restart the cycle.

This description of the WTRSD reveals the many intended uses of curriculum-associated tests. They are used for selection--to group together students with common instructional needs--and for assessment--to certify competence at a skill and to assess progress towards skill attainment. Although they are used to suggest strengths and weaknesses across objectives, they provide only uncertain diagnoses of strengths and weaknesses within objectives.

Items are generated for curriculum-associated tests in a variety of ways. The most typical way is to base items upon a list of objectives that has been compiled without a definite theory; Houghton-Mifflin publishes reading curricula (e.g., Fiesta, level 9, 1971; Serendipity, level 13, 1971) for which items have been generated in this way. A few curricula are organized around objectives which are specifically ordered; Distar, (1969), is an example. Still others have depended upon more loosely defined item generation procedures. Some of these resemble the use of a table of specifications; some are simply statements of informal guidelines. Items for Dimensions in Reading--An American Album, (1968) were generated from guidelines that can be summarized as follows: Items should measure comprehension that is more literal than inferential, and should focus on both main ideas and supporting details.

Apparently, only rarely have items of curriculum-associated tests been revised systematically; i.e., the tryout and item revision phase of test development is usually omitted. Developers of objectives-based

curricula have made some adjustments to objectives which have necessitated item modifications. These adjustments usually have been based on observations of classroom use of the instructional materials rather than on data gathered from the use of the tests. Typically, item revision has involved only minor "fine-tuning" based upon informal observation. Exceptions to this norm have occurred (e.g., the more formalized revision of exercises included in the "Guides to Individual Skill Assessment" of the WDRSD), but their occurrence has been infrequent.

The precision of curriculum-associated tests very rarely has been assessed. Only one example of precision assessment could be located: item and whole-test reliabilities were obtained for each of the WTRSD.

Little evidence could be found of attempts to ensure the validity of curriculum-associated tests. Attempts at content validation have occurred with some objective-based curricula, but the usual approach has been to rely on face validity, with an emphasis upon item appearance. Face validity typically has been determined by either the curriculum authors or external "experts."

This assessment of the characteristics of curriculum-associated tests is summarized graphically in Figure 6.

Insert Figure 6 about here.

National Assessment of Educational Progress: Reading

In the 1960's, the organization now known as the National Assessment of Educational Progress (NAEP) was established, first by private funds and eventually by federal funds, to collect and furnish information regarding

the educational achievement of the nation's children, youths, and young adults. Presently, NAEP conducts assessments in ten content areas--art, career and occupational development, citizenship, literature, mathematics, music, reading, science, social studies, and writing. This discussion deals exclusively with the second (1974-75) assessment of reading (National Assessment of Educational Progress, 1974).

Tests developed by NAEP are to be used to determine the nation's reading competence as a function of time, geographical location (e.g., region, size, and type of community), and student characteristics (e.g., age, sex, and race). For example, results from an item requiring students to read a passage and answer questions about how to serve Meow-Wow cat food suggest that 32% of 9-year-olds, 54% of 13-year-olds, and 76% of 17-year-olds could follow the instructions correctly (National Assessment of Educational Progress, 1976).

Item generation was based on objectives. Panels of reading scholars, reading educators, and lay citizens assembled to establish reading objectives, design items based on the objectives, and determine weightings of importance for each objective. All objectives were organized in a subjectively-determined taxonomy with four major headings: 1) demonstrate behavior conducive to reading, 2) demonstrate word identification skills, 3) possess skills for reading comprehension, and 4) use a variety of approaches in gathering information. The subjective nature of this organizational scheme produced many disagreements among panel members (NAEP, 1974).

After items were field-tested, "bad" items (as primarily defined by extremely high or low difficulty levels) were inspected to determine if

subtle omissions and/or ambiguities had been overlooked by panel members. These inspections, supplemented by notes from field coordinators about students' comments and reactions, provided the basis for item revision. However, NAEP personnel report that the objectives were rather "tight" by the time the items were field tested. Thus, although NAEP held a "final selection conference" to assemble the final test form, changes to items were few in number and involved only minor refinement.

NAEP's indices of precision differ from traditional indices since NAEP seeks to use the performance of carefully chosen samples of individuals to draw conclusions about the status of large, well-defined groups. Thus, precision has been determined by estimates of sampling variability around the estimates of proportions of people who can respond to items correctly. These estimates are not unlike the precision indices recommended for domain-referenced measures.

The validity of the NAEP tests appears to have been determined by the panels of scholars, educators, and lay citizens. Their discussions of the objectives and items helped ensure content and face validity.

On the basis of the above analysis, it is possible to produce a profile of the NAEP reading test. This profile appears in Figure 7.

Insert Figure 7 about here.

Conclusions

In the first three sections of this report, a model for describing and evaluating tests of reading comprehension has been developed and applied.

The model incorporates five important dimensions of test characteristics: usage, item development, item revision, assessment of precision, and validation procedures. Each dimension is presented as a scale whose left-most endpoint represents the status of the protypic standardized, norm-referenced achievement test and whose right end describes the status of the idealized domain-referenced achievement measure. With the endpoints on the five dimensions anchored in this manner, it was possible to identify, on each scale, one or more intermediate points.

The model was then employed to describe the characteristics of five major types of existing achievement tests: standardized tests (as represented in the Anchor Test Study), criterion-referenced tests (using the SOBAR series produced by SRA), individual psychoeducational tests (represented by the Illinois Test of Psycholinguistic Abilities), curriculum associated tests, and the reading test from the National Assessment of Educational Progress.

Benefits of the Model

It appears that the model constitutes a reasonable way to describe various types of reading comprehension tests. Determining where a particular test is located on each of the model dimensions was usually not a difficult task. The necessary information was either readily available from test manuals or easily obtained by phone calls to publishers and/or developers of the tests. If all test publishers would adhere to the guidelines set forth in the Standards for Educational and Psychological Tests (American Psychological Association, 1974), virtually every test

could be profiled using this model. Preliminary evidence, then, suggests that the proposed model is both viable and productive.

Implications of the Model

In developing this model, we found that in addition to its descriptive value, it has the benefit of being prescriptive. That is, the analysis of test characteristics led to the recognition that there are identifiably "good" and "not-so-good" profiles of test characteristics. In particular, a straight line running down either the left side or the right side of the set of scales in the descriptive model represents a consistent inference system. That is, item generation, test revision, assessment of precision, and validation procedures are all designed and carried out in such a way that they support the use for which the test is intended. These four characteristics are logically dependent on the intended use of the test. Tests or testing systems which are represented by lines running in a zigzag pattern across the scales suggest inconsistencies in the systems of inference they are using. Most exisiting tests show a zigzag pattern to some extent (cf. Figure 8).

Insert Figure 8 about here.

The zigzag pattern may be a signal of any one of three different situations. First, and most serious, it may signal a deep conceptual confusion, for example, an attempt to diagnose strengths and weaknesses through a testing system whose only measure of precision is classical, rank-order-of-students reliability. Second, the zigzag pattern may represent a test or testing system with multiple intended uses. Some of these systems

may be useful, but they must be carefully scrutinized by users. Third, as illustrated by the previously reviewed SOBAR system, the zigzag pattern may represent a domain-referenced testing system in early stages of empirical development. Since it is so difficult to generate large-scale, practical, domain-referenced tests in which all the rules of inference are maximally satisfied, it may often be necessary to work with approximations of this kind.

This study began as an attempt to specify the role of domain-referenced measures in assessing reading comprehension. However, domain-referenced tests are not profiled in this report. There is a simple explanation for this omission: to date, no generally available domain-referenced test of reading comprehension has been found. As indicated above, a profile for such tests would be depicted by a straight line down the right side of each scale. None of the tests examined has a profile coming very close to this. Obviously, tests intended to assess an individual's status with respect to some domain or to monitor change at frequent intervals do not yet exist with a consistent inference system supporting their use.

Two rather simple morals may be drawn from this review: (1) It is not possible to change the purposes of tests or testing from one end of the scale to the other without simultaneously building up the necessary supporting inference system. It is neither possible to use a domain-referenced test to select students without assessing its reliability for that purpose, nor possible to use a norm-referenced test for the purpose of diagnosing without assessing its reliability for that purpose. This is not yet widely recognized. (2) It is possible to have a single testing system that serves both purposes only if both the inference systems are built up. One may

select items from a well defined domain either for the purpose of rank ordering individuals or for the purpose of assessing an individual's strengths and weaknesses. But to do so one must have different sampling plans, different procedures for assessing precision, and validity corresponding to each of the purposes and plans.

It is worth re-emphasizing that virtually all considerations in test development and score interpretation are contingent on the intended uses of the test. To be specific, when the intended use of the test is to reliably rank order, there is no available approach that has either cost effectiveness or common sense advantages over a norm-referenced test (as exemplified by standardized tests of reading comprehension). Problems with this approach obviously arise when frequent assessment becomes necessary. In this case, multiple forms of the instrument are needed, and the standard procedures that can adequately ensure that one form of the test is reliable and valid must be applied to all forms of the test. Consequently, the test development procedures must be moved along the dimensions from the end primarily associated with tests that differentiate (e.g., standardized, norm-referenced tests) toward that associated with tests that measure (e.g., domain-referenced tests), and the mechanics of generalizability theory must be incorporated into the inference system.

When the primary use of tests is to certify competence, there is a suggested profile for developing them. This profile is characterized by an objectives approach to the generation of items and the establishment of rules for adjusting objectives when items need revision. The reliability and validity procedures for this type of test demand procedures which are

characteristic of domain-referenced tests. That is, it is important that repeated measures on individuals be obtained, and that the validity of the items be consistent with various constructs known to exist in that content area.

The Potential Role of Domain-Referenced Tests

In general then, what can be said about the assessement needs of reading comprehension? When tests are needed to sort students and to make relative judgments about which students can comprehend better than others, the standardized norm-referenced test will generally do the job. Currently, there are a number of good standardized tests. In addition, findings from the Anchor Test Study (Bianchini and Loret, 1974) support the notion that results from many of these tests are comparable and that forms and tests can be interchanged since they measure the same general behaviors. However, it must be emphasized that these tests cannot be used to make decisions about day to day instruction or to assess the effects of curricula and classroom procedures.

When making decisions about which curriculum package to choose, when profiling strengths and weaknesses in students' reading activities, or when deciding which among several exercises or experiences a group of students needs next, what might be described as "middle of the road" objectives-based tests are useful. The profile of such tests runs approximately down the middle of Figure 2. Their primary uses are to certify competence and to diagnose strengths and weaknesses. Such testing systems, as illustrated by SOBAR, currently have serious limitations. However, they could be improved on the basis of experience if funds were provided for long-term programs of field testing.

When problems arise which are related to classroom instructional procedures or the diagnosis of student difficulties, more refined measures are needed. In addition, basic and applied studies of factors affecting reading comprehension require sensitive and relevant outcome measures. It is especially for these kinds of situations that the use of "pure-form" domain-referenced measures is called for.

We are very aware that domain-referenced tests are costly to generate (Anderson, et al., 1978). Also, while they have a certain ease of administration and interpretability, it is difficult to balance those features against the problems of rigorously developing the tests and of using sophisticated models of reading as bases for item generation procedures. However, the high cost of developing the theories and rules which underpin the development of test-generation procedures is compensated for later on. Once a domain has been specified and the item forms developed, specific test forms can be generated routinely.

Domain-referenced tests have an important role to play in assessing reading comprehension. As a sound theoretical base for the processes of reading comprehension continues to evolve, their usefulness should grow, for it is out of sound theoretical bases that domains can be identified and elaborated.

References

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Anastasi, A. Psychological testing (4th ed.). New York: Macmillan, 1976.

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Anderson, R. C., & Faust, G. W. Educational psychology: The science of instruction and learning. New York: Dodd-Mead, 1973.

Anderson, T. H., Wardrop, J. L., Hively, W., Muller, K. E., Anderson, R. I., Hastings, C. N., & Frederiksen, J. Development and trial of a model for developing domain referenced tests of reading comprehension. Tech. Rep. No. 86. Urbana: University of Illinois, Center for the Study of Reading, May 1978.

Bianchini, J. C., & Loret, P. G. Anchor Test study final report. Berkeley, Calif.: Educational Testing Service, 1974. (ERIC Document Reproduction Service No. ED 092 601)

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Buros, O. K. Fifty years in testing: Some reminiscences, criticisms, and suggestions. Educational Researcher, 1977 (July-August), 6 (7), 9-15.

Carroll, J. A review of the Revised Illinois Test of Psycholinguistic Abilities. In O. K. Buros (Ed.), Seventh mental measurements yearbook. Highland Park, N.J.: Gryphon Press, 1972. Pp. 819-823.

Chase, C. I. A review of the Revised Illinois Test of Psycholinguistic Abilities. In O. K. Buros (Ed.), Seventh mental measurements yearbook. Highland Park, N. J.: Gryphon Press, 1972, Pp. 823-825.

Coombs, C. H. Theory and methods of social measurement. In L. Festinger & D. Katz (Eds.), Research methods in the behavioral sciences. New York: Dryden, 1953. Pp. 471-535.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington: American Council on Education, 1971. Pp. 443-507.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Dimensions in reading--An American album. Chicago: Science Research Associates, 1968.

Distar Reading I, II, III. Chicago: Science Research Associates, 1969.

Fiesta (level 9). Boston: Houghton-Mifflin, 1971.

Gronlund, N. E. Measurement and evaluation in teaching (3rd ed.). New York: Macmillan, 1976.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project (CSE Monograph Series in Evaluation, No. 1). Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Jenkins, J. R., & Pany, D. Reading comprehension in the middle grades: Instruction and research. Unpublished manuscript, University of Illinois at Urbana-Champaign, 1977.

Kirk, S. A., McCarthy, J. J., & Kirk, W. D. Examiner's manual: Illinois Test of Pscyholinguistic Abilities (rev. ed.). Urbana: University of Illinois Press, 1969.

Loevinger, J. Objective tests as instruments of psychological theory. Psychological Reports, 1957, 3, 635-694 (Monograph Supplement 9).

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Macready, G. B. The structure of domain hierarchies found within a domain referenced testing system. Educational and Psychological Measurement, 1975, 35, 583-598.

Martuza, V. R. Applying norm-referenced and criterion-referenced measurement in education. Boston: Allyn and Bacon, 1977.

McCarthy, J. J., & Kirk, S. A. The Illinois Test of Psycholinguistic Abilities. Urbana: University of Illinois Press, 1961.

McCarthy, J. J., & Olsen, J. L. Validity studies on the Illinois Test of Psycholinguistic Abilities. Madison: University of Wisconsin, 1964.

National Assessment of Educational Progress. Objectives for the second reading assessment. Denver: Education Commission of the States, 1974.

National Assessment of Educational Progress. A profile of the second reading assessment. Denver: Educational Commission of the States, 1976.

Osgood, C. E. Method and theory in experimental psychology. New York: Oxford Press, 1953.

Osgood, C. E. A behavioristic analysis: Contemporary approaches to cognition. Cambridge: Harvard University Press, 1957. (a)

Osgood, C. E. Motivational dynamics of language behavior. Nebraska Symposium on Motivation. Lincoln: University of Nebraska Press, 1957. (b)

Otto, W., & Askov, E. The Wisconsin design for reading skill development: Rationale and guidelines. Minneapolis, Minn.: National Computer Systems, 1974.

Paraskevopoulos, J. N., & Kirk, S. A. The development and psychometric characteristics of the Revised Illinois Test of Psycholinguistic Abilities. Urbana: University of Illinois Press, 1969.

Peak, H. Problems of objective observation. In L. Festinger & D. Datz (Eds.), Research methods in the behavioral sciences. New York: Dryden, 1953. Pp. 243-299.

SCORE. Geneva, Iowa: Houghton-Mifflin, 1975.

Serendipity (level 13). Boston: Houghton-Mifflin, 1971.

Short Mastery Tests: SOBAR Reading. Chicago: Science Research Asscoiates, 1976.

Wardrop, J. L. Standardized testing in the schools: Uses and roles. Monterey, Calif.: Brooks/Cole, 1976.

Table 1

Comparison of Tests Used to Differentiate

and Tests Used to Measure

	Tests Used to Differentiate	Tests Used to Measure
Major appropriate use	Differentiate (rank order) individuals, usually to sort into categories	Estimate "how much" of a given characteristic is possessed by each individual
Test characteristics	Regular, probably infrequent, administration (small number of alternate forms sufficient)	Administration as needed or "on demand," perhaps frequently (need possibly large number of alternate forms)
	Importance of reference group in test development	Importance of theory governing item development and content selection
Test development	Descriptive model (Process x Content grid), specific items are important	Predictive model (variables affecting difficulty, domain identification), item forms are important
	Try out items, compare observed with desired performance, modify to approximate desired performance	Try out items, compare observed with predicted performance, revise model to account for observation
Assessing test quality	Reliable: replicability of rank ordering, discrimination among individuals	Repeatable: consistent estimates of status (or pattern of change) within individuals
	Valid: adequacy of content representation, prediction of ranking on criterion	Valid: adequacy of construct representation, fit to model, prediction of item difficulty

Figure Captions

Figure 1. Repeatability of assessment.

Figure 2. Descriptive model for profiling test characteristics.

Figure 3. Profile of standardized achievement tests.

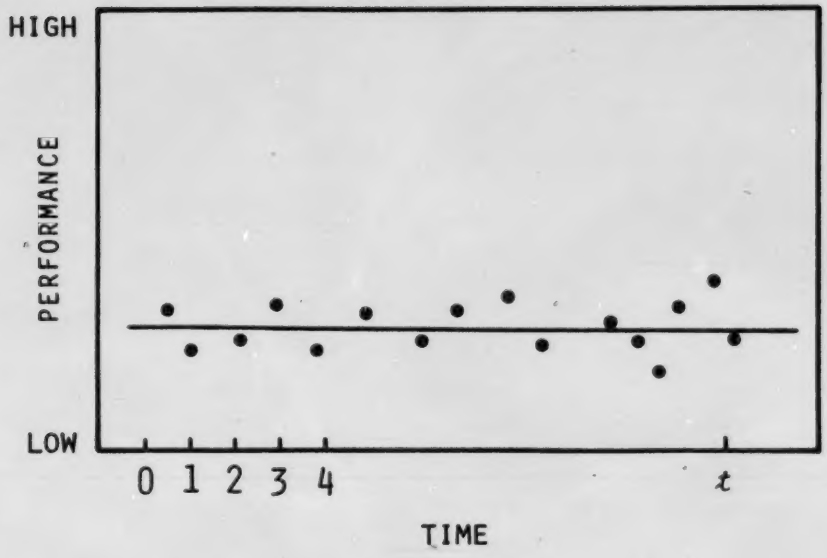
Figure 4. Profile of criterion-referenced tests.

Figure 5. Profile of psycho-educational tests (ITPA).

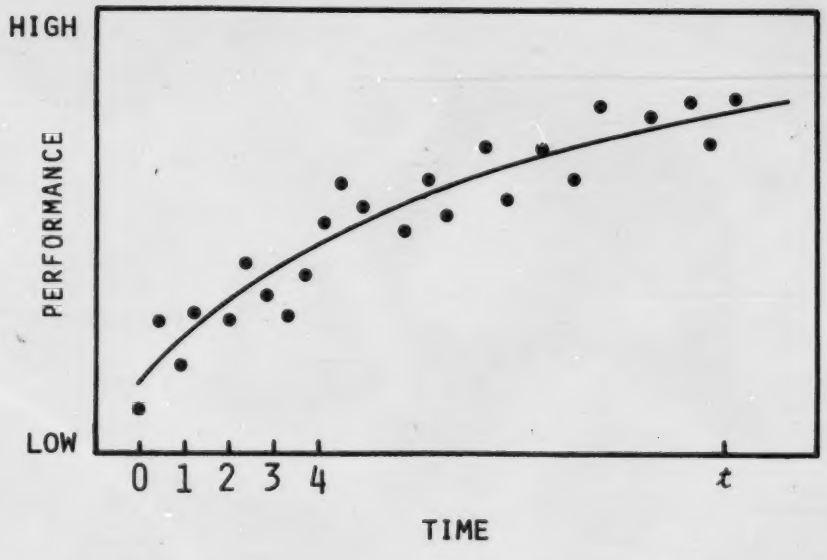
Figure 6. Profile of curriculum-associated tests.

Figure 7. Profile of state/national assessment (NAEP).

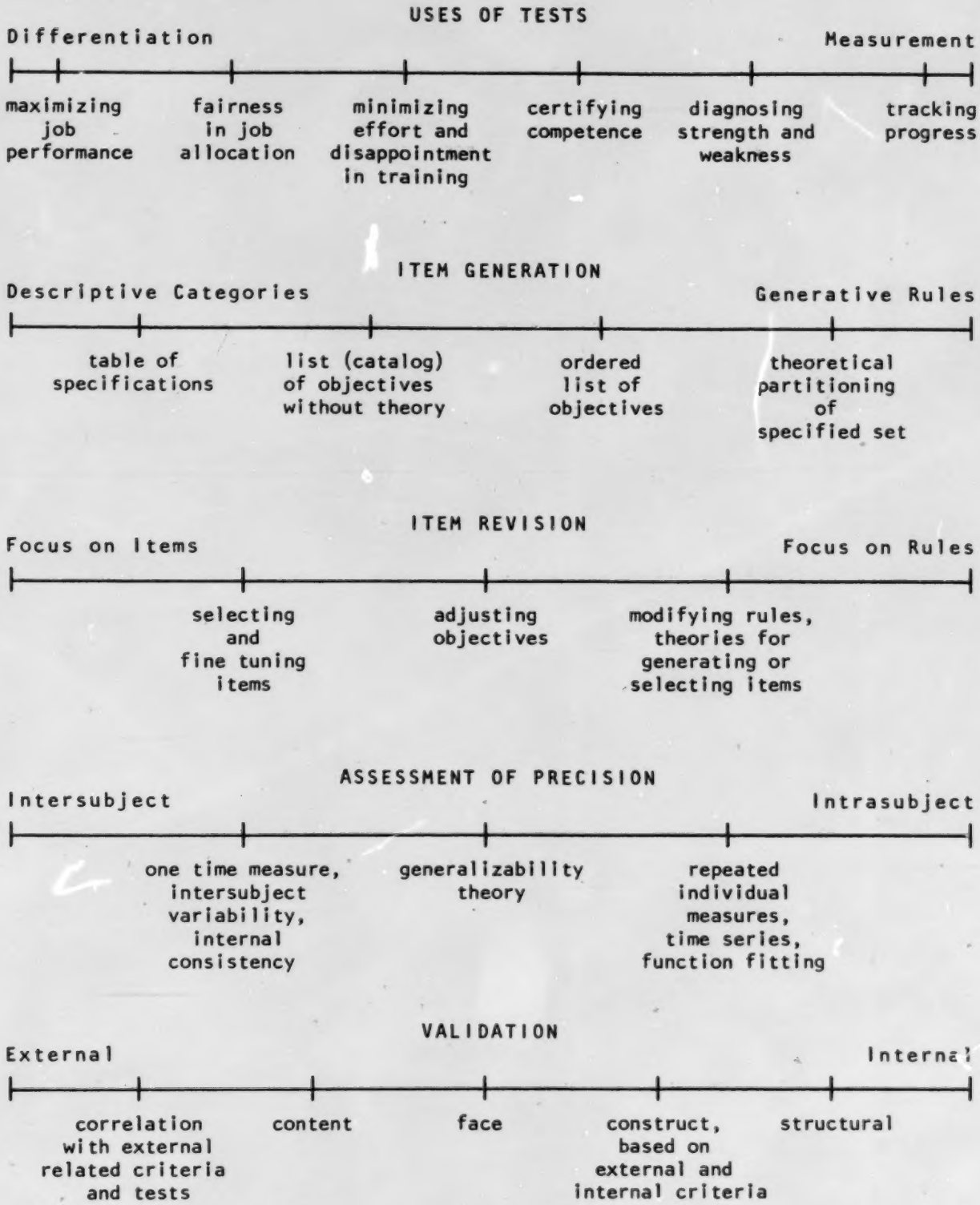
Figure 8. Comparison of test profiles.

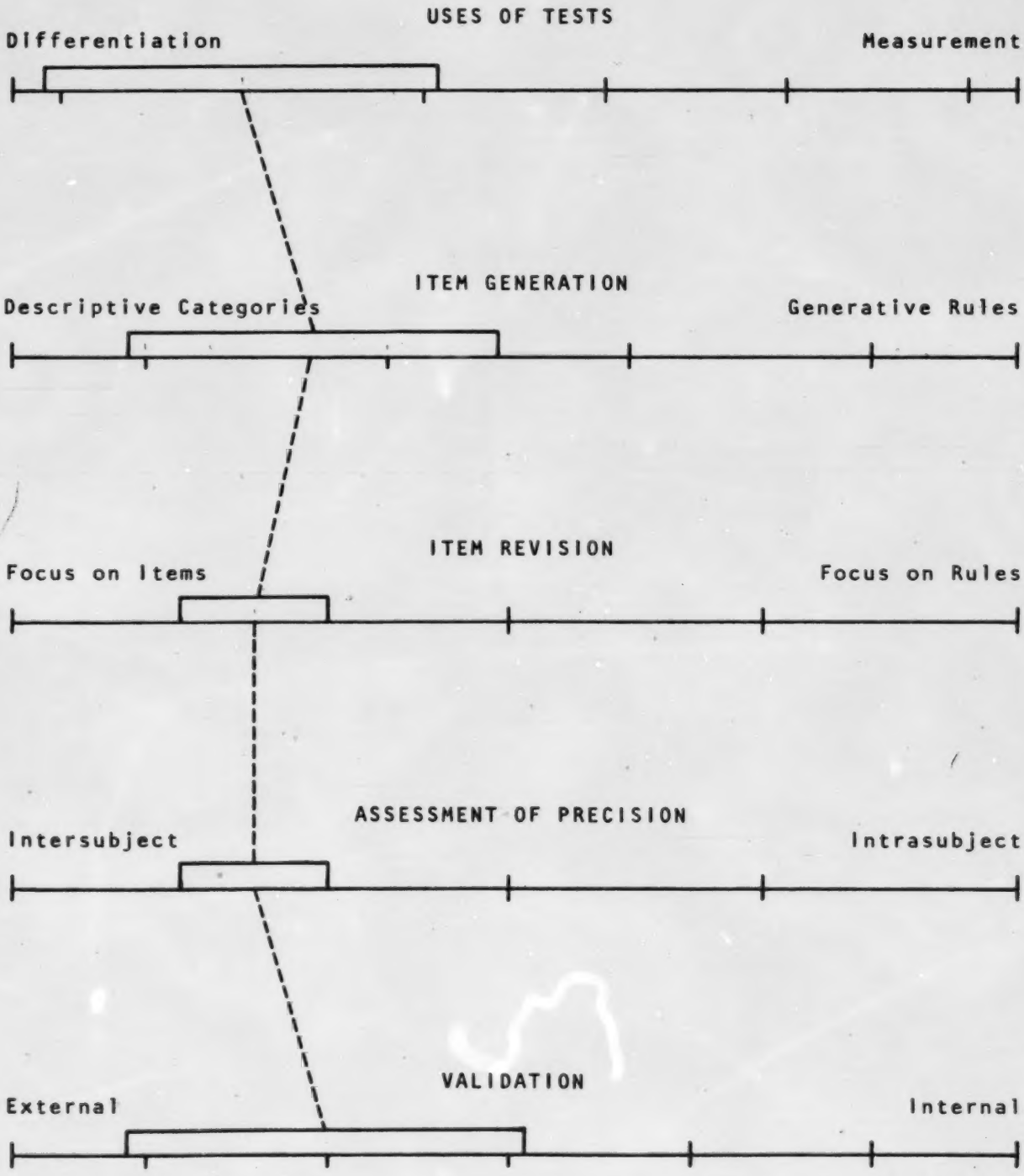


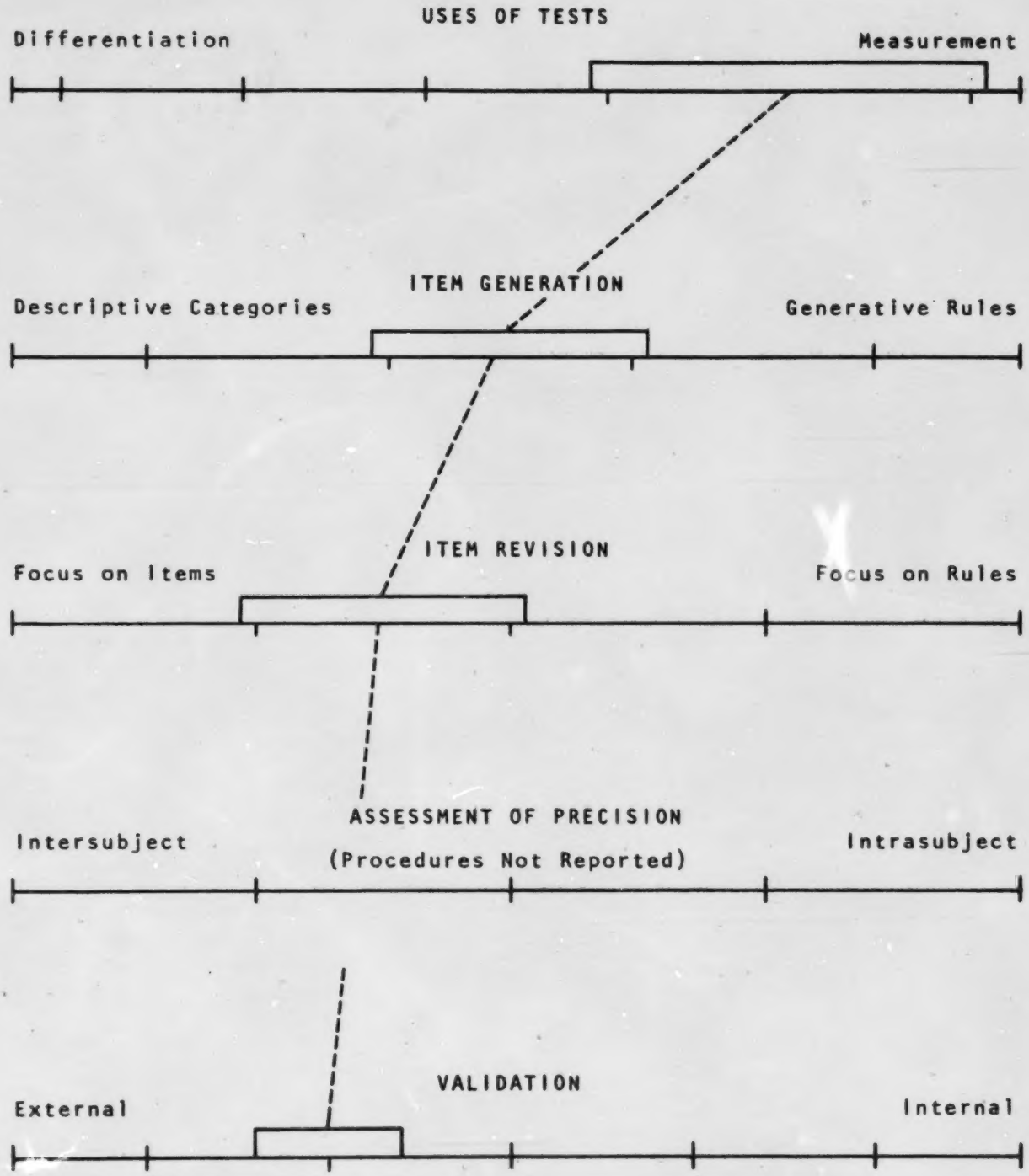
a) Stable characteristic, consistency over time

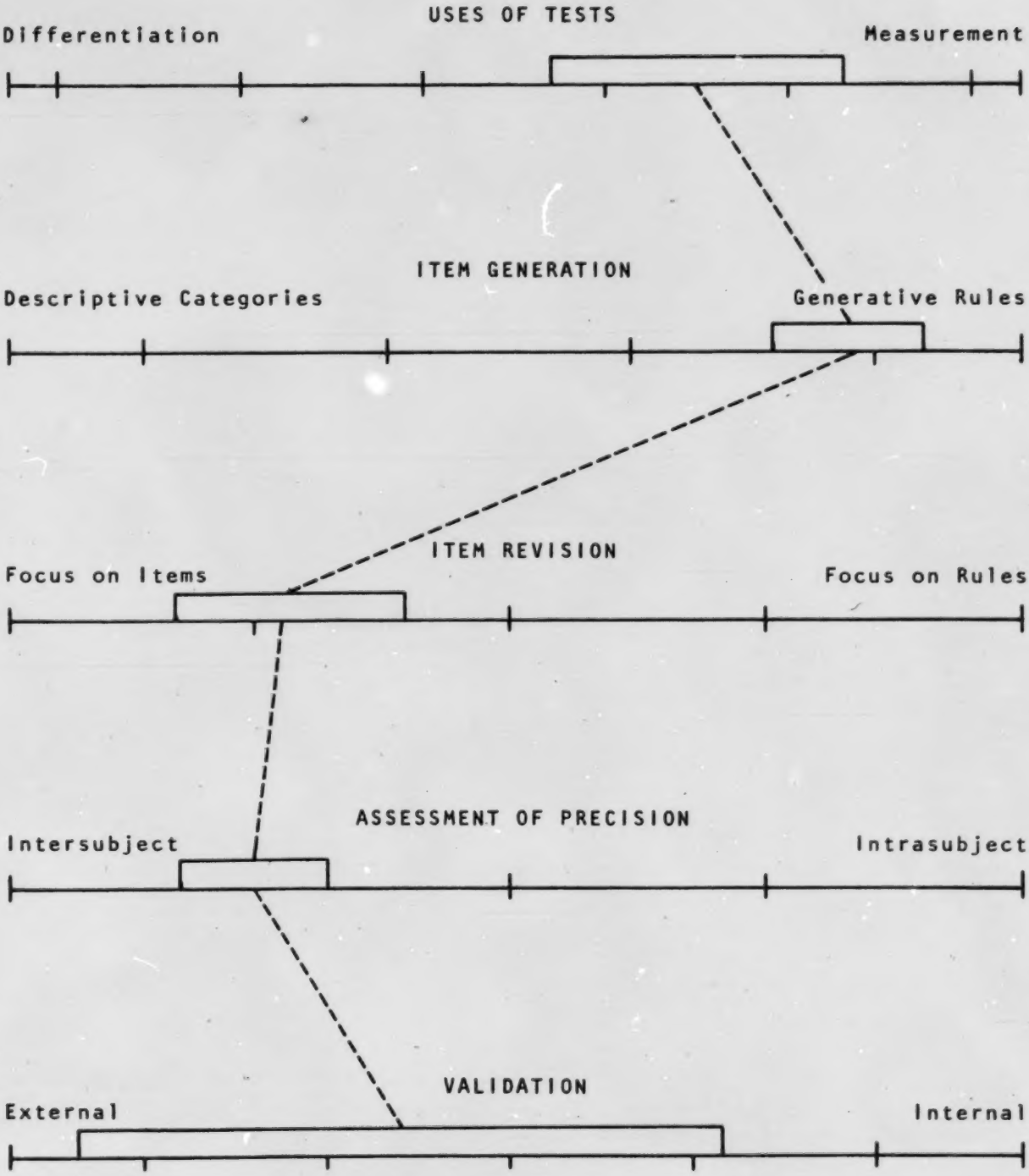


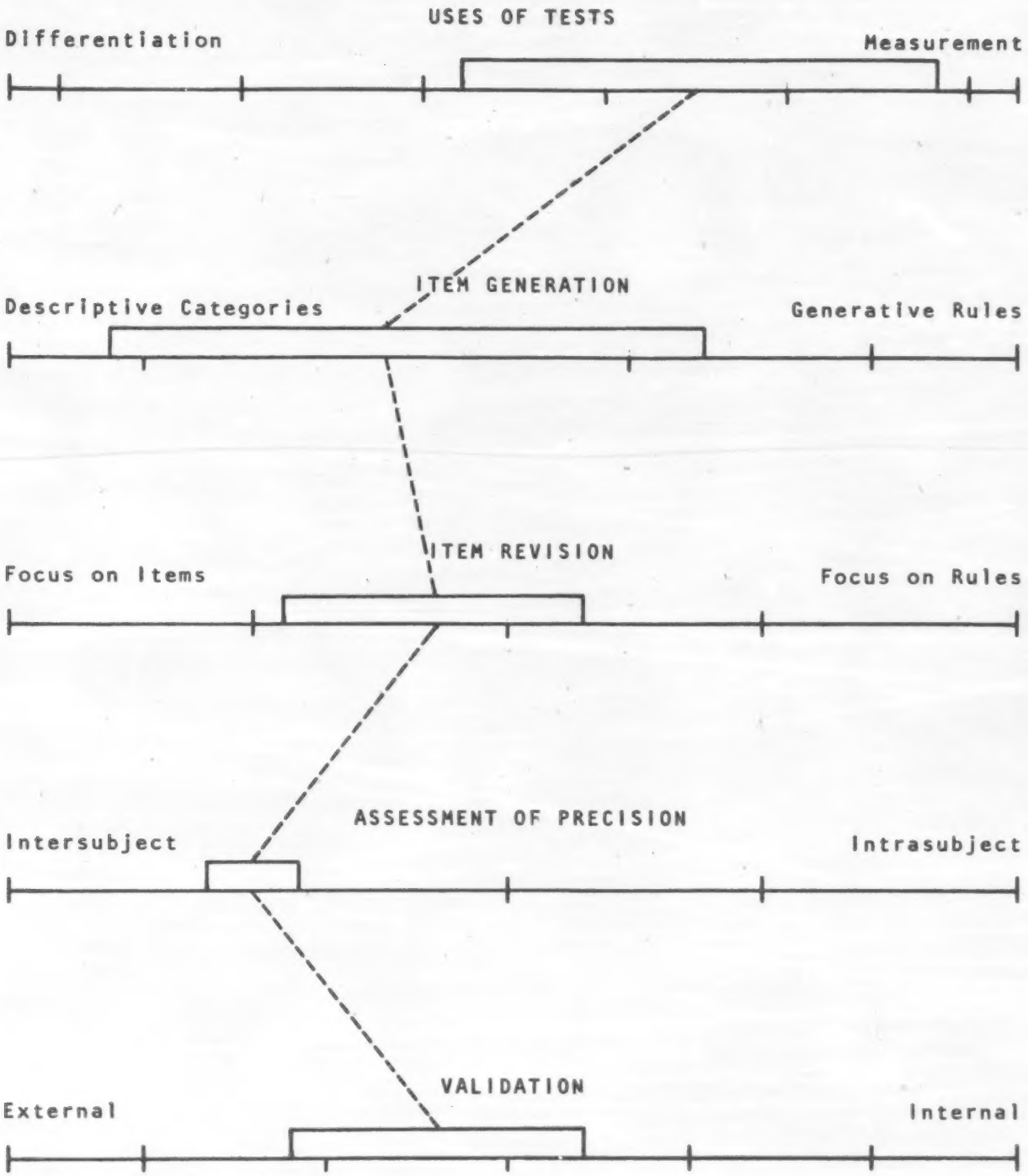
b) Developmental characteristic, changing over time

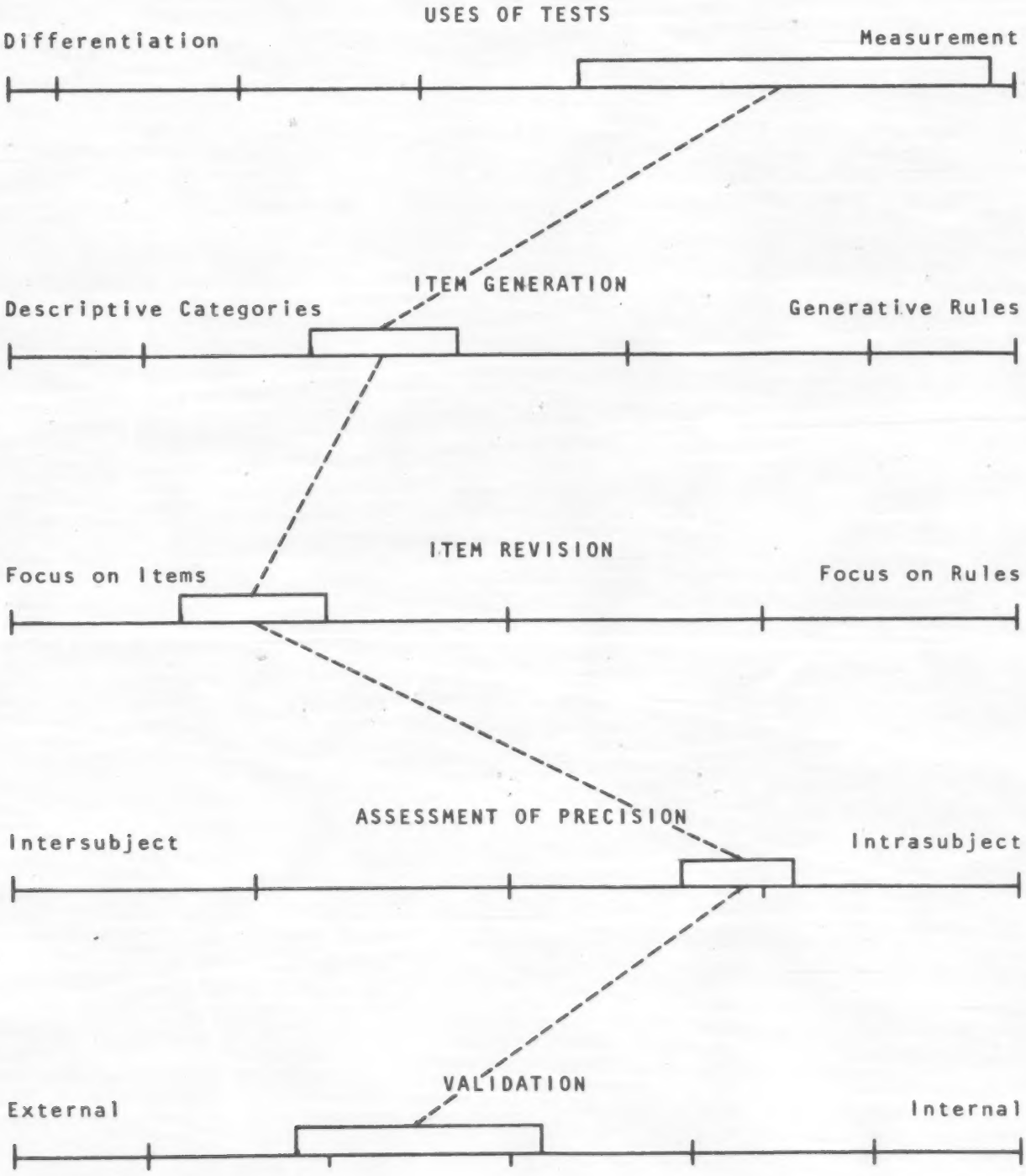

USES OF TESTS
Differentiation
Measurement
maximizing job performance
fairness in job allocation
minimizing effort and disappointment in training
certifying competence
diagnosing strength and weakness
tracking progress
ITEM GENERATION
Descriptive Categories
Generative Rules
table of specifications
list (catalog) of objectives without theory
ordered list of objectives
theoretical partitioning of specified set
ITEM REVISION
Focus on Items
Focus on Rules
selecting and fine tuning items
adjusting objectives
modifying rules, theories for generating or selecting items
ASSESSMENT OF PRECISION
Intersubject
Intrasubject
one time measure, intersubject variability, internal consistency
generalizability theory
repeated individual measures, time series, function fitting
VALIDATION
External
Internal
correlation with external related criteria and tests
content
face
construct, based on external and internal criteria
structural

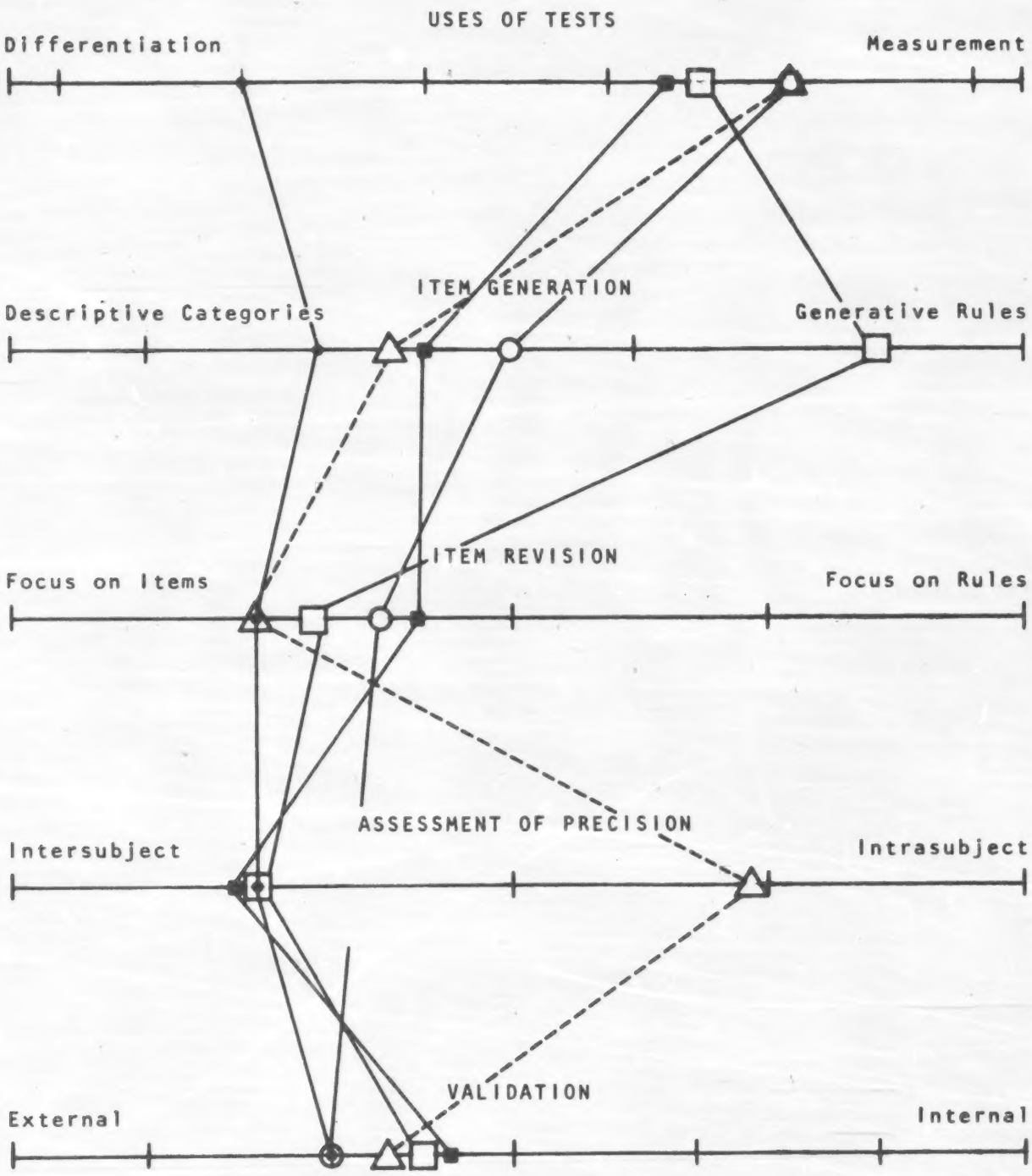

USES OF TESTS
Differentiation
Measurement
ITEM GENERATION
Descriptive Categories
Generative Rules
ITEM REVISION
Focus on Items
Focus on Rules
ASSESSMENT OF PRECISION
Intersubject
Intrasubject
VALIDATION
External
Internal


USES OF TESTS
Differentiation
Measurement
ITEM GENERATION
Descriptive Categories
Generative Rules
ITEM REVISION
Focus on Items
Focus on Rules
ASSESSMENT OF PRECISION
(Procedures Not Reported)
Intersubject
Intrasubject
VALIDATION
External
Internal


USES OF TESTS
Differentiation
Measurement
ITEM GENERATION
Descriptive Categories
Generative Rules
ITEM REVISION
Focus on Items
Focus on Rules
ASSESSMENT OF PRECISION
Intersubject
Intrasubject
VALIDATION
External
Internal


USES OF TESTS
Differentiation
Measurement
ITEM GENERATION
Descriptive Categories
Generative Rules
ITEM REVISION
Focus on Items
Focus on Rules
ASSESSMENT OF PRECISION
Intersubject
Intrasubject
VALIDATION
External
Internal


USES OF TESTS
Differentiation
Measurement
ITEM GENERATION
Descriptive Categories
Generative Rules
ITEM REVISION
Focus on Items
Focus on Rules
ASSESSMENT OF PRECISION
Intersubject
Intrasubject
VALIDATION
External
Internal



Standardized
Criterion-referenced
Psycho-educational
Curriculum-associated
National Assessment of Educational Progress (NAEP)

CENTER FOR THE STUDY OF READING

READING EDUCATION REPORTS

No. 1: Durkin, D. Comprehension Instruction--Where Are You?, October 1977. (ERIC Document Reproduction Service No. ED 146 566, 14p., HC-$1.67, MF-$.83)

No. 2: Asher, S. R. Sex Differences in Reading Achievement, October 1977. (ERIC Document Reproduction Service No. ED 146 567, 30p., HC-$2.06, MF-$.83)

No. 3: Adams, M. J., Anderson, R. C., & Durkin, D. Beginning Reading: Theory and Practice, November 1977.

No. 4: Jenkins, J. R., & Pany, D. Teaching Reading Comprehension in the Middle Grades, January 1978.

No. 5: Bruce, B. What Makes a Good Story?, June 1978.

No. 6: Anderson, T. H. Another Look at the Self-Questioning Study Technique, September 1978.

CENTER FOR THE STUDY OF READING

TECHNICAL REPORTS

* Available only through ERIC

*No. 1: Halff, H. M. Graphical Evaluation of Hierarchical Clustering Schemes, October 1975. (ERIC Document Reproduction Service No. ED 134 926, 11p., HC-$1.67, MF-$.83)

*No. 2: Spiro, R. J. Inferential Reconstruction in Memory for Connected Discourse, October 1975. (ERIC Document Reproduction Service No. ED 136 187, 81p., HC-$4.67, MF-$.83)

*No. 3: Goetz, E. T. Sentences in Lists and in Connected Discourse, November 1975. (ERIC Document Reproduction Service No. ED 134 927, 75p., HC-$3.50, MF-$.83)

*No. 4: Alessi, S. M., Anderson, T. H., & Biddle, W. B. Hardware and Software Considerations in Computer Based Course Management, November 1975. (ERIC Document Reproduction Service No. ED 134 928, 21p., HC-$1.67, MF-$.83)

*No. 5: Schallert, D. L. Improving Memory for Prose: The Relationship Between Depth of Processing and Context, November 1975. (ERIC Document Reproduction Service No. ED 134 929, 37p., HC-$2.06, MF-$.83)

*No. 6: Anderson, R. C., Goetz, E. T., Pichert, J. W., & Halff, H. M. Two Faces of the Conceptual Peg Hypothesis, January 1976. (ERIC Document Reproduction Service No. ED 134 930, 29p., HC-$2.06, MF-$.83)

*No. 7: Ortony, A. Names, Descriptions, and Pragmatics, February 1976. (ERIC Document Reproduction Service No. ED 134 931, 25p., HC-$1.67, MF-$.83)

No. 8: Mason, J. M. Questioning the Notion of Independent Processing Stages in Reading, February 1976. (Journal of Educational Psychology, 1977, 69, 288-297)

*No. 9: Siegel, M. A. Teacher Behaviors and Curriculum Packages: Implications for Research and Teacher Education, April 1976. (ERIC Document Reproduction Service No. ED 134 932, 42p., HC-$2.06, MF-$.83)

*No. 10: Anderson, R. C., Pichert, J. W., Goetz, E. T., Schallert, D. L., Stevens, K. V., & Trollip, S. R. Instantiation of General Terms, March 1976. (ERIC Document Reproduction Service No. ED 134 933, 30p., HC-$2.06, MF-$.83)

*No. 11: Armbruster, B. B. Learning Principles from Prose: A Cognitive Approach Based on Schema Theory, July 1976. (ERIC Document Reproduction Service No. ED 134 934, 48p., HC-$2.06, MF-$.83)

*No. 12: Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. Frameworks for Comprehending Discourse, July 1976. (ERIC Document Reproduction Service No. ED 134 935, 33p., HC-$2.06, MF-$.83)

No. 13: Rubin, A. D., Bruce, B. C., & Brown, J. S. A Process-Oriented Language for Describing Aspects of Reading Comprehension, November 1976. (ERIC Document Reproduction Service No. ED 136 188, 41p., HC-$2.06, MF-$.83)

No. 14: Pichert, J. W., & Anderson, R. C. Taking Different Perspectives on a Story, November 1976. (ERIC Document Reproduction Service No. ED 134 936, 30p., HC-$2.06, MF-$.83)

No. 15: Schwartz, R. M. Strategic Processes in Beginning Reading, November 1976. (ERIC Document Reproduction Service No. ED 134 937, 19p., HC-$1.67, MF-$.83)

No. 16: Jenkins, J. R., & Pany, D. Curriculum Biases in Reading Achievement Tests, November 1976. (ERIC Document Reproduction Service No. ED 134 938, 24p., HC-$1.67, MF-$.83)

No. 17: Asher, S. R., Hymel, S., & Wigfield, A. Children's Comprehension of High- and Low-Interest Material and a Comparison of Two Cloze Scoring Methods, November 1976. (ERIC Document Reproduction Service No. ED 134 939, 32p., HC-$2.06, MF-$.83)

No. 18: Brown, A. L., Smiley, S. S., Day, J. D., Townsend, M. A. R., & Lawton, S. C. Intrusion of a Thematic Idea in Children's Comprehension and Retention of Stories, December 1976. (ERIC Document Reproduction Service No. ED 136 189, 39p., HC-$2.06, MF-$.83)

No. 19: Kleiman, G. M. The Prelinguistic Cognitive Basis of Children's Communicative Intentions, February 1977. (ERIC Document Reproduction Service No. ED 134 940, 51p., HC-$3.50, MF-$.83)

No. 20: Kleiman, G. M. The Effect of Previous Context on Reading Individual Words, February 1977. (ERIC Document Reproduction Service No. ED 134 941, 76p., HC-$4.67, MF-$.83)

No. 21: Kane, J. H., & Anderson, R. C. Depth of Processing and Interference Effects in the Learning and Remembering of Sentences, February 1977. (ERIC Document Reproduction Service No. ED 134 942, 29p., HC-$2.06, MF-$.83)

No. 22: Brown, A. L., & Campione, J. C. Memory Strategies in Learning: Training Children to Study Strategically, March 1977. (ERIC Document Reproduction Service No. ED 136 234, 54p., HC-$3.50, MF-$.83)

No. 23: Smiley, S. S., Oakley, D. D., Worthen, D., Campione, J. C., & Brown, A. L. Recall of Thematically Relevant Material by Adolescent Good and Poor Readers as a Function of Written Versus Oral Presentation, March 1977. (ERIC Document Reproduction Service No. ED 136 235, 23p., HC-$1.67, MF-$.83)

No. 24: Anderson, R. C., Spiro, R. J., & Anderson, M. C. Schemata as Scaffolding for the Representation of Information in Connected Discourse, March 1977. (ERIC Document Reproduction Service No. ED 136 236, 18p., HC-$1.67, MF-$.83)

No. 25: Pany, D., & Jenkins, J. R. Learning Word Meanings: A Comparison of Instructional Procedures and Effects on Measures of Reading Comprehension with Learning Disabled Students, March 1977. (ERIC Document Reproduction Service No. ED 136 237, 34p., HC-$2.06, MF-$.83)

No. 26: Armbruster, B. B., Stevens, R. J., & Rosenshine, B. Analyzing Content Coverage and Emphasis: A Study of Three Curricula and Two Tests, March 1977. (ERIC Document Reproduction Service No. ED 136 238, 22p., HC-$1.67, MF-$.83)

No. 27: Ortony, A., Reynolds, R. E., & Arter, J. A. Metaphor: Theoretical and Empirical Research, March 1977. ERIC Document Reproduction Service No. ED 137 752, 63p., HC-$3.50, MF-$.83)

No. 28: Ortony, A. Remembering and Understanding Jabberwocky and Small-Talk, March 1977. (ERIC Document Reproduction Service No. ED 137 753, 36p., HC-$2.06, MF-$.83)

No. 29: Schallert, D. L., Kleiman, G. M., & Rubin, A. D. Analysis of Differences Between Oral and Written Language, April 1977. (ERIC Document Reproduction Service No. ED 144 038, 33p., HC-$2.06, MF-$.83)

No. 30: Goetz, E. T., & Osborn, J. Procedures for Sampling Texts and Tasks in Kindergarten through Eighth Grade, April 1977. (ERIC Document Reproduction Service No. ED 146 565, 80p., HC-$4.67, MF-$.83)

No. 31: Nash-Webber, B. Anaphora: A Cross-Disciplinary Survey, April 1977. (ERIC Document Reproduction Service No. ED 144 039, 43p., HC-$2.06, MF-$.83)

No. 32: Adams, M. J., & Collins, A. A Schema-Theoretic View of Reading Comprehension, April 1977. (ERIC Document Reproduction Service No. ED 146 565, 80p., HC-$4.67, MF-$.83)

No. 33: Huggins, A. W. F. Syntactic Aspects of Reading Comprehension, April 1977. (ERIC Document Reproduction Service No. ED 142 972, 68p., HC-$3.50, MF-$.83)

No. 34: Bruce, B. C. Plans and Social Actions, April 1977. (ERIC Document Reproduction Service No. ED 149 328, 45p., HC-$2.06, MF-$.83)

No. 35: Rubin, A. D. Comprehension Processes in Oral and Written Language, April 1977. (ERIC Document Reproduction Service No. ED 150 550, 61p., HC-$3.50, MF-$.83)

No. 36: Nash-Webber, B., & Reiter, R. Anaphora and Logical Form: On Formal Meaning Representations for Natural Language, April 1977. (ERIC Document Reproduction Service No. ED 142 973, 42p., HC-$2.06, MF-$.83)

No. 37: Adams, M. J. Failures to Comprehend and Levels of Processing in Reading, April 1977.

No. 38: Woods, W. A. Multiple Theory Formation in High-Level Perception, April 1977. (ERIC Document Reproduction Service No. ED 144 020, 58p., HC-$3.50, MF-$.83)

No. 40: Collins, A., Brown, J. S., & Larkin, K. M. Inference in Text Understanding, December 1977. (ERIC Document Reproduction Service No. ED 150 547, 48p., HC-$2.06, MF-$.83)

No. 41: Anderson, R. C., & Pichert, J. W. Recall of Previously Unrecallable Information Following a Shift in Perspective, April 1977. (ERIC Document Reproduction Service No. ED 142 974, 37p., HC-$2.06, MF-$.83)

No. 42: Mason, J., Osborn, J., & Rosenshine, B. A Consideration of Skill Hierarchy Approaches to the Teaching of Reading, December 1977. (ERIC Document Reproduction Service No. ED 150 549, 176p., HC-$10.03, MF-$.83)

No. 43: Collins, A., Brown, A. L., Morgan, J. L., & Brewer, W. F. The Analysis of Reading Tasks and Texts, April 1977.

No. 44: McClure, E. Aspects of Code-Switching in the Discourse of Bilingual Mexican-American Children, April 1977. (ERIC Document Reproduction Service No. ED 142 975, 39p., HC-$2.06, MF-$.83)

No. 45: Schwartz, R. M. Relation of Context Utilization and Orthographic Automaticity in Word Identification, May 1977.

No. 46: Anderson, R. C., Stevens, K. C., Shifrin, Z., & Osborn, J. Instantiation of Word Meanings in Children, May 1977. (ERIC Document Reproduction Service No. ED 142 976, 22p., HC-$1.67, MF-$.83)

No. 47: Brown, A. L. Knowing When, Where, and How to Remember: A Problem of Metacognition, June 1977. (ERIC Document Reproduction Service No. ED 146 562, 152p., HC-$8.69, MF-$.83)

No. 48: Brown, A. L., & DeLoache, J. S. Skills, Plans, and Self-Regulation, July 1977. (ERIC Document Reproduction Service No. ED 144 040, 66p., HC-$3.50, MF-$.83)

No. 49: Goetz, E. T. Inferences in the Comprehension of and Memory for Text, July 1977. (ERIC Document Reproduction Service No. ED 150 548, 97p., HC-$4.67, MF-$.83)

No. 50: Anderson, R. C. Schema-Directed Processes in Language Comprehension, July 1977. (ERIC Document Reproduction Service No. ED 142 977, 33p., HC-$2.06, MF-$.83)

No. 51: Brown, A. L. Theories of Memory and the Problems of Development: Activity, Growth, and Knowledge, July 1977. (ERIC Document Reproduction Service No. ED 144 041, 59p., HC-$3.50, MF-$.83)

No. 52: Morgan, J. L. Two Types of Convention in Indirect Speech Acts, July 1977.

No. 53: Brown, A. L., Smiley, S. S., & Lawton, S. C. The Effects of Experience on the Selection of Suitable Retrieval Cues for Studying from Prose Passages, July 1977. (ERIC Document Reproduction Service No. ED 144 042, 30p., HC-$2.06, MF-$.83)

No. 54: Fleisher, L. S., & Jenkins, J. R. Effects of Contextualized and Decontextualized Practice Conditions on Word Recognition, July 1977. (ERIC Document Reproduction Service No. ED 144 043, 37p., HC-$2.06, MF-$.83)

No. 55: Jenkins, J. R., & Larson, K. Evaluating Error Correction Procedures for Oral Reading, June 1978.

No. 56: Anderson, T. H., Standiford, S. N., & Alessi, S. M. Computer Assisted Problem Solving in an Introductory Statistics Course, August 1977. (ERIC Document Reproduction Service No. ED 146 563, 26p., HC-$2.06, MF-$.83)

No. 57: Barnitz, J. Interrelationship of Orthography and Phonological Structure in Learning to Read, August 1977. (ERIC Document Reproduction Service No. ED 150 546, 62p., HC-$3.50, MF-$.83)

No. 58: Mason, J. M. The Role of Strategy in Reading in the Mentally Retarded, September 1977.

No. 59: Mason, J. M. Reading Readiness: A Definition and Skills Hierarchy from Preschoolers' Developing Conceptions of Print, September 1977.

No. 60: Spiro, R. J., & Esposito, J. J. Superficial Processing of Explicit Inferences in Text, December 1977. (ERIC Document Reproduction Service No. ED 150 545, 27p., HC-$2.06, MF-$.83)

No. 65: Brewer, W. F. Memory for the Pragmatic Implications of Sentences, October 1977. (ERIC Document Reproduction Service No. ED 146 564, 27p., HC-$2.06, MF-$.83)

No. 66: Brown, A. L., & Smiley, S. S. The Development of Strategies for Studying Prose Passages, October 1977.

No. 68: Stein, N. L., & Nezworski, T. The Effects of Organization and Instructional Set on Story Memory, January 1978. (ERIC Document Reproduction Service No. ED 149 327, 41p., HC-$2.06, MF-$.83)

No. 69: Stein, N. L. How Children Understand Stories: A Developmental Analysis, March 1978.

No. 76: Thieman, T. J., & Brown, A. L. The Effects of Semantic and Formal Similarity on Recognition Memory for Sentences in Children, November 1977. (ERIC Document Reproduction Service No. ED 150 551, 26p., HC-$2.06, MF-$.83)

No. 77: Nash-Webber, B. L. Inference in an Approach to Discourse Anaphora, January 1978. (ERIC Document Reproduction Service No. ED 150 552, 30p., HC-$2.06, MF-$.83)

No. 78: Gentner, D. On Relational Meaning: The Acquisition of Verb Meaning, December 1977. (ERIC Document Reproduction Service No. ED 149 325, 46p., HC-$2.06, MF-$.83)

No. 79: Royer, J. M. Theories of Learning Transfer, January 1978. (ERIC Document Reproduction Service No. ED 149 326, 55p., HC-$3.50, MF-$.83)

No. 80: Arter, J. A., & Jenkins, J. R. Differential Diagnosis-Prescriptive Teaching: A Critical Appraisal, January 1978. (ERIC Document Reproduction Service No. ED 150 578, 104p., HC-$6.01, MF-$.83)

No. 81: Shoben, E. J. Choosing a Model of Sentence Picture Comparisons: A Reply to Catlin and Jones, February 1978. (ERIC Document Reproduction Service No. ED 150 577, 30p., HC-$2.06, MF-$.83)

No. 82: Steffensen, M. S. Bereiter and Engelmann Reconsidered: The Evidence from Children Acquiring Black English Vernacular, March 1978.

No. 83: Reynolds, R. E., Standiford, S. N., & Anderson, R. C. Distribution of Reading Time when Questions are Asked about a Restricted Category of Text Information, April 1978.

No. 84: Baker, L. Processing Temporal Relationships in Simple Stories: Effects of Input Sequence, April 1978.

No. 85: Mason, J. M., Knisely, E., & Kendall, J. Effects of Polysemous Words on Sentence Comprehension, May 1978.

No. 86: Anderson, T. H., Wardrop, J. L., Hively, W., Muller, K. E., Anderson, R. I., Hastings, C. N., & Frederiksen, J. Development and Trial of a Model for Developing Domain Referenced Tests of Reading Comprehension, May 1978.

No. 87: André, M. E. D. A., & Anderson, T. H. The Development and Evaluation of a Self-Questioning Study Technique, June 1978.

No. 88: Bruce, B., & Newman, D. Interacting Plans, June 1978.

No. 89: Bruce, B., Collins, A., Rubin, A. D., & Gentner, D. A Cognitive Science Approach to Writing, June 1978.

No. 90: Asher, S. T. Referential Communication, June 1978.

No. 91: Royer, J. M., & Cunningham, D. J. On the Theory and Measurement of Reading Comprehension, June 1978.

No. 92: Mason, J. M., & Kendall, J. R. Facilitating Reading Comprehension through Text Structure Manipulation, June 1978.

No. 93: Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. Interpreting Metaphors and Idioms: Some Effects of Context on Comprehension, July 1978.

No. 94: Brown, A. L., Campione, J. C., & Barclay, C. R. Training Self-Checking Routines for Estimating Test Readiness: Generalization from List Learning to Prose Recall, July 1978.

No. 95: Reichman, R. Conversational Coherency, July 1978.

No. 96: Wigfield, A., & Asher, S. R. Age Differences in Children's Referential Communication Performance: An Investigation of Task Effects, July 1978.

No. 97: Steffensen, M. S., Jogdeo, C., & Anderson, R. C. A Cross-Cultural Perspective on Reading Comprehension, July 1978.

No. 98: Green, G. M. Discourse Functions of Inversion Construction, July 1978.

No. 99: Asher, S. R. Influence of Topic Interest on Black Children and White Children's Reading Comprehension, July 1978.

No. 100: Jenkins, J. R., Pany, D., & Schreck, J. Vocabulary and Reading Comprehension: Instructional Effects, August 1978.

No. 101: Shoben, E. J., Rips, L. J., & Smith, E. E. Issues in Semantic Memory: A Response to Glass and Holyoak, August 1978.

No. 102: Baker, L., & Stein, N. L. The Development of Prose Comprehension Skills, September 1978.

No. 103: Fleisher, L. S., Jenkins, J. R., & Pany, D. Effects on Poor Readers' Comprehension of Training in Rapid Decoding, September 1978.

No. 104: Anderson, T. H. Study Skills and Learning Strategies, September 1978.

No. 105: Ortony, A. Beyond Literal Similarity, October 1978.

No. 106: Durkin, D. What Classroom Observations Reveal about Reading Comprehension Instruction, October 1978.

No. 107: Adams, M. J. Models of Word Recognition, October 1978.

No. 108: Reder, L. M. Comprehension and Retention of Prose: A Literature Review, November 1978.

No. 109: Wardrop, J. L., Anderson, T. H., Hively, W., Anderson, R. I., Hastings, C. N., & Muller, K. E. A Framework for Analyzing Reading Test Characteristics, December 1978.